ESTADÍSTICA I (ADE): TEORÍA Y EJERCICIOS

Victoria Alea Riera

Ernest Jiménez Garrido

Carme Muñoz Vaquer

Núria Viladomiu Canela

Curso 2015/16

PRESENTACIÓN

Esta publicación electrónica va dirigida a los alumnos de Estadística I del grado de Administración y Dirección de Empresas de la Universidad de Barcelona.

Su principal objetivo es proporcionar un material de trabajo que facilite la dinámica de las clases presenciales y ayude al alumno en su planificación del estudio.

Cada capítulo de la publicación corresponde a un tema del programa. Se inicia con un breve resumen de los conceptos básicos e imprescindibles. A continuación se proporciona un conjunto de ejercicios que permitirán trabajar los conceptos de la asignatura.

Estos ejercicios están ordenados siguiendo el orden utilizado en la programación de las sesiones presenciales y también el grado de dificultad de los ejercicios.

Este material es fruto de años de experiencia práctica de las autoras en la asignatura Estadística y se ha constatado que proporciona a los alumnos un instrumento útil para el seguimiento de la asignatura.

Barcelona, julio 2015

Tema 1. CONCEPTO Y CONTENIDO DE LA ESTADÍSTICA

Objeto de la estadística
Estadística descriptiva e inferencia estadística
Población y muestra
Datos. Clasificación y escalas de medida
Instalación del programa R-Commander

La ESTADÍSTICA da respuesta a preguntas como son:

- ¿Cuál será la proporción de electores que votarán a un partido determinado en unas elecciones municipales?
- ¿Cuál es el porcentaje de unidades defectuosas con que opera determinado proceso de producción?
- ¿Cuál es precio de los spots publicitarios en televisión?
- ¿Han variado en los últimos 5 años los alquileres de los locales comerciales en la ciudad de Barcelona?
- ¿Cómo repercute sobre la demanda de un producto un incremento en su precio?
- ¿Cómo se relacionan la tasa de inflación y la tasa de paro de un país?

La estadística permite reducir la incertidumbre en el proceso de toma de decisiones en el ámbito empresarial, económico, político, etc.

El proceso estadístico comienza identificando el grupo cuyo comportamiento se quiere describir. Este grupo recibe el nombre de POBLACIÓN. La población estadística está formada no sólo por personas, sino por cualquier tipo de objetos o entidades sobre los cuales pueda observarse alguna característica.

Por ejemplo, se quiere averiguar la proporción de electores de Badalona que votarán a un determinado candidato en las próximas elecciones municipales. En este caso la población está formada por todos los habitantes censados en Badalona con capacidad de voto.

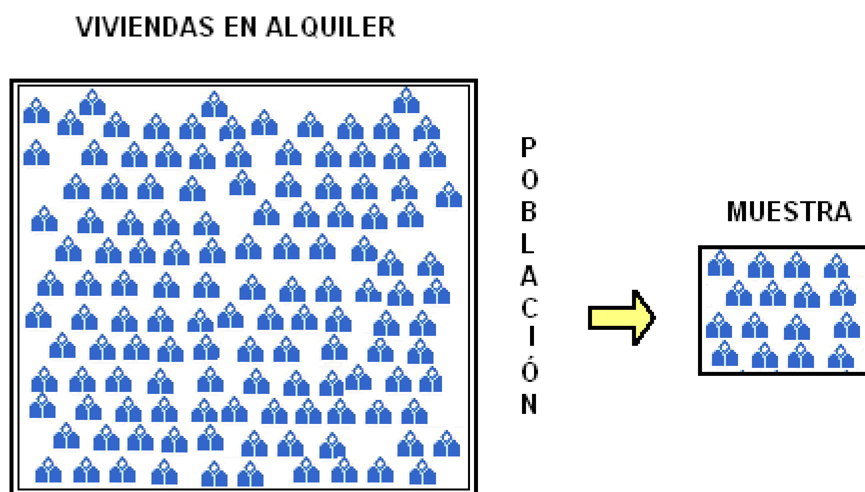
Un fabricante quiere calcular el porcentaje de unidades defectuosas con que opera su proceso de producción. En este caso la población la constituyen todas las unidades fabricadas mientras el proceso se mantenga en su actual estado. En este caso el número de elementos de la población es teóricamente infinito.

Resultados académicos de la última prueba de acceso a la universidad en Andalucía. En este caso la población son los estudiantes que han realizado esta prueba en esa Comunidad.

Importe del alquiler en agosto de los apartamentos para 4 personas en las localidades costeras de la comarca de La Selva. En este caso la población son los apartamentos que reúnen las características de localización y tamaño indicadas.

Es básico diferenciar si la información de que se dispone corresponde a toda la población objeto de estudio o está limitada a una parte de la misma o MUESTRA.

Cuando la población contiene un número infinito o muy grande de elementos es imposible observar la característica de interés sobre cada uno de ellos. En este caso el análisis estadístico se basa en la observación de un subconjunto de la población que recibe el nombre de muestra.

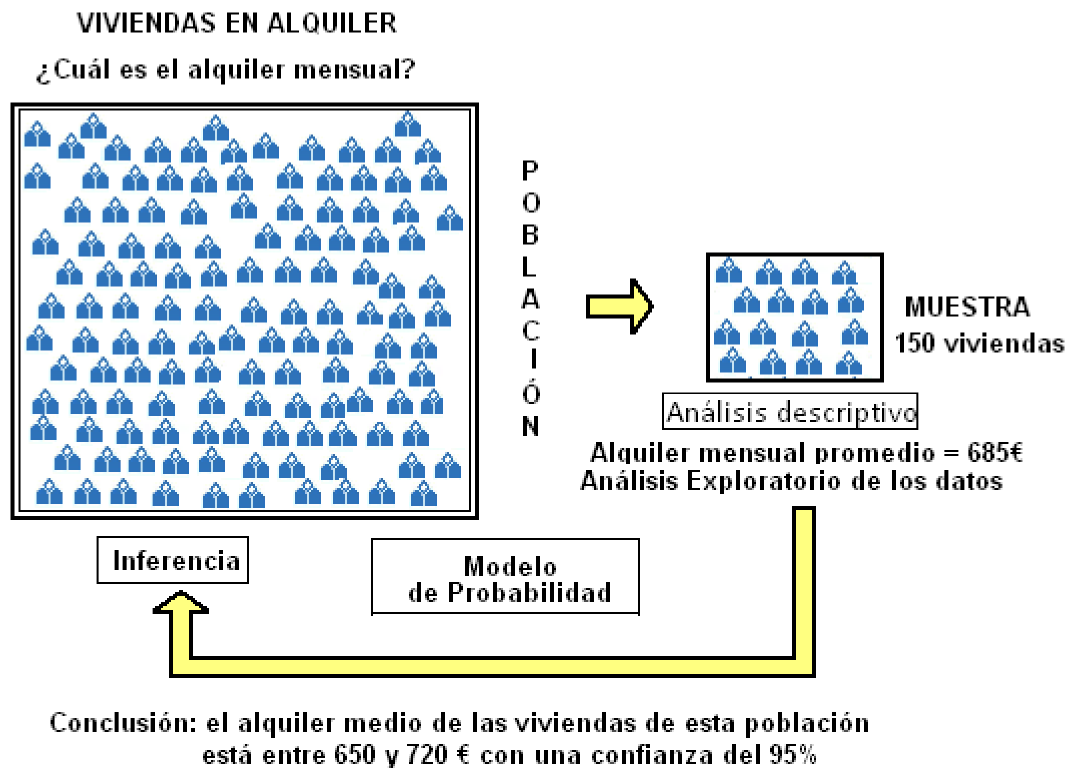


La ESTADÍSTICA DESCRIPTIVA y el ANÁLISIS EXPLORATORIO DE DATOS recogen un conjunto de técnicas que permiten el resumen de los datos y la descripción de la muestra.

El análisis exploratorio permite, además, identificar patrones de comportamiento de los datos y formular hipótesis sobre la población que podrán ser validadas mediante las técnicas del análisis confirmatorio que utiliza los métodos de la INFERENCIA ESTADÍSTICA.

Dada la naturaleza limitada de la información muestral, al inferir (inducir) el comportamiento de la población a partir de la descripción de la muestra, es necesario evaluar la fiabilidad de los resultados en términos probabilísticos.

La TEORÍA DE LA PROBABILIDAD permite calcular el margen de error con el que puede aceptarse el modelo matemático o teórico de comportamiento propuesto para la población.



DATOS

La observación de la característica de interés en la muestra proporciona los DATOS. Los datos pueden consistir en un conjunto de valores numéricos o modalidades. Por ejemplo, si se sondea a la población de electores de Badalona sobre su intención de votar a determinado candidato los datos presentan dos modalidades: SI/NO. En el caso de que se analicen los resultados académicos de los estudiantes de Andalucía, los datos serán valores numéricos, de 0 a 10. En el caso de que se analice el importe del alquiler de las viviendas de una localidad, los datos son valores numéricos en Euros.

Las VARIABLES son las características de los individuos que se quieren estudiar y pueden tomar distintas modalidades o valores.

Los DATOS son el conjunto de observaciones de una o más características obtenidas de una población o de una muestra.

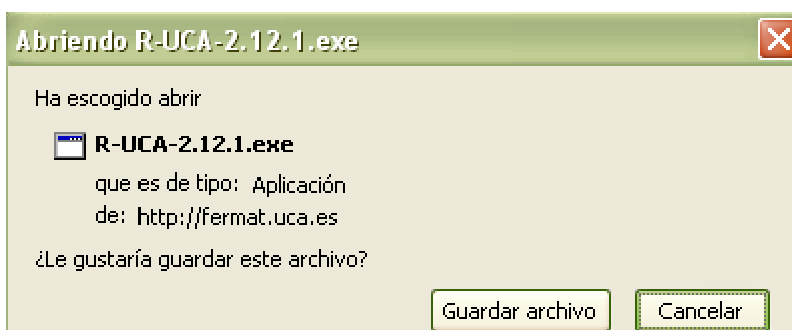
Es importante distinguir entre los distintos tipos de datos con los que podemos tratar. Sus diferencias determinan la selección y aplicación de las técnicas estadísticas

- 1. Naturaleza del fenómeno
  - CUALITATIVOS Atributos
    - NOMINALES
    - ORDINALES
  - CUANTITATIVOS Variables
    - DISCRETAS
    - CONTINUAS
- 2. Referidas al Tiempo
  - Corte Transversal
  - Series Temporales
  - Datos de Panel
- 3. Número de Características
  - Unidimensionales
  - Multidimensionales (Bidimensionales)

INSTALACIÓN DE R COMMANDER

Para instalar la versión de R y R COMMANDER adaptada por la Universidad de Cádiz es preciso:

- Descargar el programa que se encuentra en la página: http://knuth.uca.es
- Elegir la opción: Paquete R UCA
- Descargar la última versión desde el servidor 2. Aparecerá el siguiente cuadro de diálogo[1].



- Almacenar el fichero R-UCA-xxx.exe en el subdirectorio de las bajadas de internet del ordenador clicando el botón Guardar archivo.
- Ejecutar el programa de instalación clicando sobre R-UCA-xxx.exe. Aparecerá el siguiente cuadro de diálogo:



- Activar la opción Ejecutar
  Finalizada con éxito la instalación en la barra de Programas aparecerá el acceso a R.

---

[1] Si se navega con Internet Explorer el cuadro de diálogo es ligeramente distinto y ofrece la posibilidad de ejecutar sin guardar.

Tema 2. DISTRIBUCIÓN DE FRECUENCIAS Y REPRESENTACIÓN GRÁFICA

Tabla de frecuencias simple: variable discreta
Diagrama de barras y de frecuencias acumuladas
Tabla de frecuencias con valores agrupados: variable continua
Histograma y polígonos de frecuencias
Análisis exploratorio de datos: diagrama de tallo y hojas (Stem and Leaf)

TABLA DE FRECUENCIAS

Recoge de forma resumida el conjunto de datos resultantes de la observación de una variable en un colectivo o muestra de n individuos.

Elementos de una tabla de frecuencias

1. Tabla de frecuencias con los valores de la variable sin agrupar:

Xi	ni	fi	Ni	Fi
x1 x2 .... xi ... xk	n1 n2 ... ni ... nk	f1 f2 ... fi ... fk	N1 N2 ... Ni ... Nk=n	F1 F2 ... Fi ... Fk=1
	n	1		

Interpretación de las columnas de la tabla:

- Xi, valores de la variable, recoge cada uno de los valores observados de X ordenados de menor a mayor.
- ni, frecuencia absoluta del valor xi, es el número de elementos de la muestra para los que X = xi.

  $\sum_{i=1}^{k} n_i = n$ La suma de todas las frecuencias absolutas es igual a n.
- fi, frecuencia relativa, es la proporción en tanto por uno de elementos para los que X = xi. fi = ni/n

  $\sum_{i=1}^{k} f_i = 1$ La suma de todas las frecuencias relativas es igual a 1.

Si se multiplican las frecuencias relativas por 100 se obtienen los correspondientes porcentajes.

- Ni, frecuencia absoluta acumulada hasta el valor xi, es el número de elementos para los que X ≤ xi.

  $Ni = n1 + n2 + \ldots + ni\text{-}1 + ni = \sum_{j=1}^{i} n_j$

- Fi, frecuencia relativa acumulada hasta xi, es la proporción de elementos para los que X ≤ xi.

  $Fi = f1 + f2 + \ldots + fi\text{-}1 + fi = \sum_{j=1}^{i} f_j$

  Si las frecuencias relativas acumuladas se multiplican por 100 se obtiene los porcentajes acumulados.

2. Tabla de frecuencias con los valores de la variable agrupados en intervalos.

Li-1-Li	Xi (ci)	ni	fi	Ni	Fi
L0-L1 L1-L2 …. Li-1-Li Lk-1-Lk	x1 x2 …. xi … xk	ni n2 …. ni .... nk	f1 f2 .... fi .... fk	N1 N2 …. Ni .... Nk=n	F1 F2 …. Fi .... Fk=1
		n	1		

Interpretación de las columnas de la tabla

- Li-1-Li recoge todos los intervalos o clases en los que se agrupan los valores de la variable; Li-1 y Li son los límites inferior y superior del intervalo i-ésimo. Los intervalos, por omisión, se establecen abiertos en el límite inferior y cerrados en el superior.
- ai, amplitud del intervalo i-ésimo, es ai = Li – Li-i.
- xi, marca de clase o punto medio del intervalo, es el valor que representa al intervalo en el análisis descriptivo, xi = $\frac{L_{i-1} + L_i}{2}$.
- ni, frecuencia absoluta del intervalo, es el número total de elementos para los que el valor de X está dentro del intervalo i-ésimo. $\sum_{i=1}^{k} n_i = n$
- fi, frecuencia relativa del intervalo, es la proporción en tanto por uno de elementos para los que X está dentro del intervalo i-ésimo, fi = ni/n.

Si se multiplican las frecuencias relativas por 100 se obtienen los correspondientes porcentajes.

$$\sum_{i=1}^{k} f_i = 1$$

- Ni, frecuencia absoluta acumulada hasta el intervalo i-ésimo, es el número de elementos para los que X ≤ Li.

  $Ni = n1 + n2 + \ldots + ni\text{-}1 + ni = \sum_{j=1}^{i} n_j$

- Fi, frecuencia relativa acumulada hasta el intervalo i-ésimo, es la proporción de elementos para los que X ≤ Li

  $Fi = f1 + f2 + \ldots + fi\text{-}1 + fi = \sum_{j=1}^{i} f_j$

  Si las frecuencias relativas acumuladas se multiplican por 100 se obtiene los porcentajes acumulados.

REPRESENTACIONES GRÁFICAS

La representación gráfica de los datos constituye un instrumento de gran utilidad ya que proporciona una imagen que permite:

- Captar de manera sencilla y rápida aspectos relevantes de la distribución de frecuencias,
- Mejorar la comprensión del fenómeno que se analiza,
- Detectar la presencia de errores en los datos.

Diagrama de barras

Para elaborar este gráfico se sitúan las categorías o valores en el eje de abscisas y en el de ordenadas las frecuencias absolutas o relativas. Sobre la marca correspondiente a cada categoría o valor se alza una barra perpendicular al eje de abscisas de altura igual a su frecuencia.

- El perfil del diagrama es el mismo si se representan las frecuencias absolutas o las frecuencias relativas.
- El criterio de orden de las categorías (datos cualitativos) más adecuado es el de mayor a menor frecuencia, mientras que el de los valores (datos cuantitativos) es de menor a mayor valor de X.
- Este gráfico permite visualizar rápidamente las categorías o valores más o menos frecuentes.

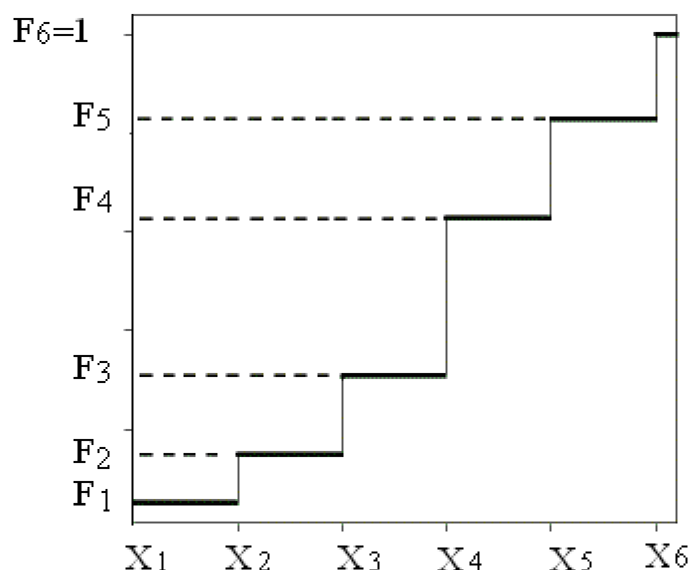
Diagrama de Escalera

El diagrama en escalera se utiliza para representar las distribuciones de frecuencias absolutas o relativas acumuladas correspondientes a una variable discreta que toma pocos valores diferentes.

Para construir el diagrama se sitúan en el eje de abscisas los valores de la variable y en el de ordenadas las frecuencias acumuladas. Se marca los puntos de coordenadas (xi, Ni) o (xi, Fi) según se quiera representar las frecuencias absolutas o relativas. Desde cada uno de estos puntos se traza una recta paralela al eje de abscisas hasta el valor siguiente de X, es decir, hasta el punto (xi+1, Ni), dado que entre dos valores consecutivos no hay acumulación de frecuencia. Los puntos extremos de las líneas horizontales se unen con líneas verticales dando al diagrama el aspecto de escalera.

- El máximo que alcanza el gráfico es n si se representan las frecuencias absolutas acumuladas o 1 si se representan las frecuencias relativas acumuladas.
- La altura de los escalones es la frecuencia absoluta o relativa de cada valor xi.

Por ejemplo, para una variable X que toma únicamente los valores x1, x2, …, x6, el diagrama de frecuencias relativas acumuladas podría ser:



Histograma

Se construye colocando en el eje de abscisas los intervalos en los que se agrupan los valores de la variable. Sobre cada intervalo se dibuja un rectángulo cuya área debe ser igual o proporcional a su frecuencia.

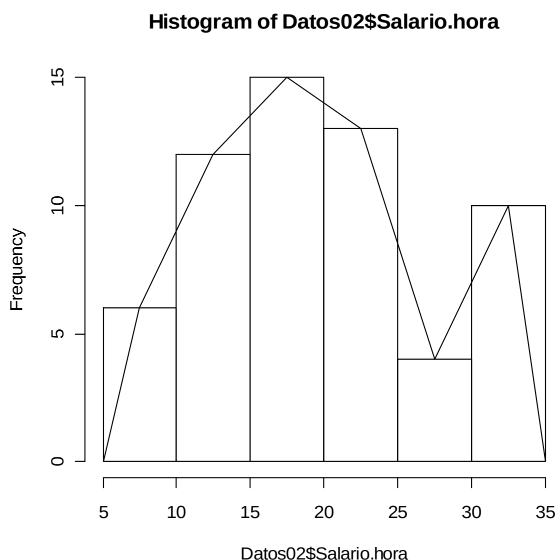
- Si todos los intervalos son de igual amplitud, por comodidad, se dibujan los rectángulos con alturas iguales a las frecuencias. En este caso, el área de los rectángulos será proporcional a la frecuencia.
- Si la amplitud del intervalo es variable se debe calcular su densidad o altura de los rectángulos. En este caso, el área de los rectángulos será igual a la frecuencia del intervalo. *Densidad = altura =* $\frac{frecuencia}{amplitud}$

En el histograma:

- Las áreas y no las alturas de los rectángulos son las que representan las frecuencias.
- La altura de los rectángulos informa sobre la densidad o concentración de observaciones en el intervalo.
- El área total es igual o proporcional al tamaño de la muestra.
- Si se representan las frecuencias relativas, el área total es igual o proporcional a 1.
- El perfil del histograma depende de la elección del número y la amplitud de los intervalos.
- El perfil del histograma es el mismo tanto si se representa la distribución de frecuencias absolutas como la de frecuencias relativas.

Polígono de Frecuencias

Es un gráfico que sintetiza el perfil del histograma y suele presentarse superpuesto a éste.



El polígono se traza señalando las marcas de clase en el lado superior de los rectángulos del histograma. Se unen estos puntos de coordenadas (x_i, n_i) (marca de clase, frecuencia absoluta o relativa) con trazo continuo y se cierra el polígono prolongándolo en sus extremos hasta cortar el eje de abscisas en los puntos situados en las marcas de clase de dos hipotéticos intervalos trazados antes que el primero y después del último.

Aunque el histograma proporciona una representación sencilla y eficaz, el polígono de frecuencias, en algunas situaciones, presenta ventajas. Dos de las razones son:

- Es más fácil comparar polígonos de varias distribuciones superponiéndolos.
- La curva suavizada del polígono sugiere de forma más clara el posible modelo de probabilidad adecuado para describir el comportamiento de la población.

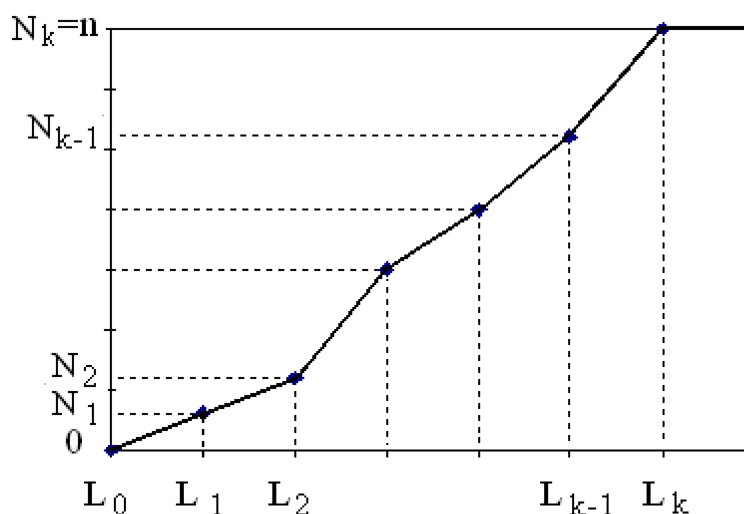
Polígono de Frecuencias acumuladas

El polígono de frecuencias acumuladas u ojiva se utiliza para representar las distribuciones de frecuencias absolutas o relativas acumuladas correspondiente a una variable cuya distribución se ha tabulado agrupando los valores en intervalos por ser de naturaleza continua o discreta que toma muchos valores diferentes.

El polígono se construye situando en el eje de abscisas los límites de los intervalos definidos en la tabla y en el de ordenadas las frecuencias acumuladas. Se señalan los puntos correspondientes a los límites superiores y sus frecuencias acumuladas, (Li, Ni) o (Li, Fi), y con trazo continuo se unen dichos puntos, empezando por el punto (L0, 0) (límite inferior del primer intervalo, frecuencia acumulada 0) y acabando en el punto (Lk, n) o (Lk, 1) (límite superior del último intervalo, frecuencia total acumulada.

Al realizar el gráfico, dado que ya no se dispone de las observaciones correspondientes a cada intervalo, se supone que éstas se reparten uniformemente en el intervalo, por lo tanto, la frecuencia se acumula de forma lineal.

Por ejemplo, para una variable X se ha tabulado en K intervalos, el diagrama de frecuencias absolutas acumuladas podría ser:



Este tipo de gráfico es adecuado si se quiere:

- Localizar valores que acumulan una determinada frecuencia. Se fija la frecuencia acumulada en el eje vertical y se localiza el valor correspondiente en el eje horizontal.
- Obtener el número o el porcentaje de observaciones con "menos que" o "más que" un valor determinado. Se fija el valor en el eje horizontal y en el eje vertical se halla la frecuencia acumulada.
- Identificar el modelo de distribución poblacional o teórica asociado a la muestra analizada superponiendo los gráficos.

Diagrama Stem-and leaf (Gráfico de tallo y hojas)

El diagrama de tallo y hojas es una técnica para presentar datos cuantitativos en formato gráfico.

Esta técnica proporciona simultáneamente:

- La ordenación de los datos. Todas las observaciones quedan ordenadas de menor a mayor, lo que facilitará la localización de algunas medidas de síntesis como son la mediana y los cuantiles.
- La tabulación de los datos. Cada tallo define un intervalo cerrado por la izquierda y abierto por la derecha equivalente al intervalo de la tabla de frecuencias con valores agrupados.
- La representación gráfica de la distribución. El perfil del gráfico es similar al histograma que se obtendría de su correspondiente tabla de frecuencias.

Al igual que el histograma, mediante el diagrama de tallo y hojas se visualizan diferentes rasgos de la distribución como son:

- Rango de los valores (dispersión)
- Localización de valores centrales
- Identificación de valores muy o poco frecuentes
- Saltos (gaps) o lagunas
- Valores anómalos o extremos notablemente desviados del conjunto
- Asimetría y forma.

Comparándolo con el histograma presenta las siguientes ventajas:

- No condensa la información. Se puede seguir reconociendo los elementos de la muestra con una mínima pérdida de información.
- Facilita la localización de los cuantiles.
- Informa de la existencia de valores outliers y los identifica.

Para construir este diagrama:

- Se divide cada valor observado en dos partes: hoja y tallo. Para ello, se fija la posición del dígito que se tomará como hoja (…, décimas, unidades, decenas, centenas, …) y los tallos quedan determinados por los dígitos que quedan a la izquierda de dicha posición.

- Se anotan en columna los tallos desde el menor hasta el mayor de forma sucesiva sin omisiones. Los tallos deben ser consecutivos y abarcar todo el recorrido de la variable.
- A la derecha de cada tallo se anotan de forma ordenada (de menor a mayor) sus hojas.
- En el encabezado o pie del gráfico es importante indicar las unidades de la hoja (o del tallo) para poder recuperar las observaciones en las unidades originales.
- Si el número de observaciones es excesivamente grande, es conveniente que cada hoja represente a un número determinado de elementos con el mismo tallo y hoja, debiéndose indicar en el diagrama.
- Se puede completar el diagrama con las frecuencias simples o acumuladas anotándolas a la izquierda de los tallos que se obtiene sumando las hojas correspondientes a cada tallo
- En el encabezado se acostumbra a indicar el tamaño de la muestra, n, que es el número total de hojas.
- En la parte superior e inferior del diagrama se anotan los outliers o valores anómalos si los hay.

Ejemplo. Supongamos que las edades de un colectivo formado por 45 trabajadores son los siguientes: 32, 32, 32, 34, 34, 35, 35, 35 ,35, 37 ,37, 37 ,37, 38 ,39, 40, 40, 41, 42, 42, 42, 42, 42, 42, 42, 43, 43, 43, 43, 43, 45, 45, 45, 45, 45, 47, 47, 48, 49, 49, 50, 50, 51, 51, 51. El gráfico de tallo y hojas de esta muestra podría ser cualquiera de los dos que siguen:

n_i	Tallo	Hojas
15	3	222445555777789
25	4	0012222222333335555577899
5	5	00111

Unidades de las hojas: 1 3|2 representa 32

Dada la poca variación que presentan los tallos es conveniente subdividirlos en partes iguales. Como las hojas toman los valores enteros de 0 a 9 únicamente se pueden subdividir los tallos en 2 o 5 partes para que sean todas ellas iguales.

Si se subdividen en 2 partes, a la primera le corresponderán las hojas del 0 al 4 y a la segunda del 5 al 9.

n_i	Tallo	Hojas
5	3	22244
10	3	5555777789
15	4	001222222233333
10	4	5555577899
5	5	00111

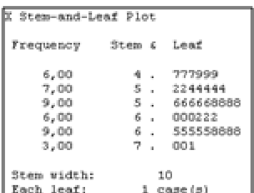
Unidades de las hojas: 1 3|2 representa 32

En el diagrama de tallo y hojas anterior se observa que la distribución es poco dispersa, los valores centrales están alrededor del 42, no presenta saltos o lagunas ni valores extremos y es simétrica.

TABULACIÓN Y REPRESENTACIÓN GRÁFICA

CUANTITATIVOS

- EXPLORATORIO
  - STEM AND LEAF



```
X Stem-and-Leaf Plot

Frequency    Stem &  Leaf

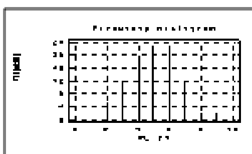
    6,00        4 .  777999
    7,00        5 .  2244444
    9,00        5 .  666668888
    6,00        6 .  000222
    9,00        6 .  555558888
    3,00        7 .  001

Stem width:         10
Each leaf:       1 case(s)
```

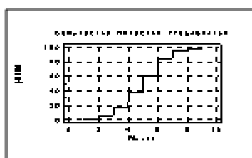
- DISTR. SIMPLES

Xi	ni	fi	Ni	Fi
x1	n1	f1	N1	F1
x2	n2	f2	N2	F2
...	...	...	...	...
xi	ni	fi	Ni	Fi
...	...	...	...	...
xk	nk	fk	N	1
	N	1		

  - BARRAS



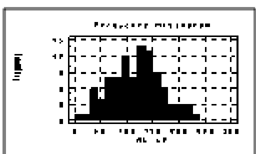
  - DIAGRAMA EN ESCALERA



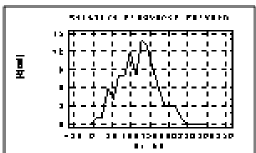
- DISTR. AGRUPADAS

Li-1-Li	ci	ni	fi	Ni	Fi
L0-L1	c1	n1	f1	N1	F1
L1-L2	c2	n2	f2	N2	F2
...	...	...	...	...	...
Li-1-Li	ci	ni	fi	Ni	Fi
...	...	...	...	...	...
Lk-1-Lk	ck	nk	fk	N	1
		N	1		

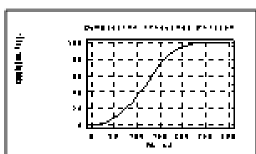
  - HISTOGRAMA



  - POLÍGONO DE FREC SIMPLE

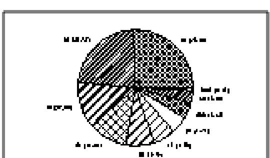


  - POLÍGONO DE FREC ACUMULADAS

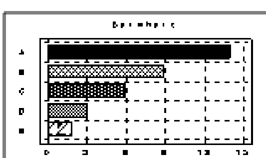


CUALITATIVOS

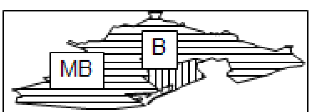
- DIAG. DE SECTORES



- DIAG. COLUMNAS O BARRAS



- CARTOGRAMAS
- PICTOGRAMAS



Actividades

Actividad 2_1

La base de datos Ejercicio21.rda procede de una encuesta realizada a 125 alumnos.

Los primeros 25 casos son:

	Genero	Edad	Miembro	Trabaja	Ingreso_men	Gasto_ocio	Lugar_resid	Medio_trans	Tiempo_viaje	Nota_acceso	Asig_matric	Asig_aprob
1	M	27	3	1	920	70	2	2	20	7.62	10	6
2	M	21	2	0		20	2	2	45	6.49	10	8
3	M	21	4	1	300	80	1			5.32	10	6
4	M	18	2	0		20	1	6	50	6.3	10	6
5	H	23	3	0		60	5	3	35	5.5	10	5
6	M	20	4	0		30	1	2	30	5.5	10	5
7	M	29	1	1	1070	50	5	4	35	6.39	10	8
8	M	19	2	1	1050	50	1	2	45	6.57	10	7
9	M	19	3	0			1	1	25	6.59	10	4
10	M	19	3	1	250	65	1	1	20	8.6	10	6
11	M	22	3	1	1150	20	2	1	20	5.95	10	8
12	M	52	4	1	1150	50	2	6	35	7.02	10	7
13	H	18	3	0		15	5	3	40	5.17	10	4
14	M	25	3	0		30	4	3	45	8.01	10	4
15	M	18	3	0		30	1	6	55	7.16	10	4
16	H	21	2	0		15	1	2	30	5.62	10	5
17	M	20	2	0		12	1	5	20	5.69	10	7
18	M	18	2	1	290	30	1	1	40	6.19	10	5
19	M	20	3	0		10	5	4	55	5.4	10	7
20	H	28	4	0		20	2	2	40	6.68	10	6
21	H	21	1	1	750	70	2	2	50	7.55	10	5
22	H	22	4	1	200	50	1	6	60	6.48	10	6
23	M	19	6	0		15	2	1	15	5.98	10	5
24	H	21	3	0		30				5.13	10	6
25	M	19	3	1	200	40	4	1	45	6.63	10	4

Las características observadas son:

Genero: H = Hombre M = Mujer
Edad
Miembros: Número de miembros de la unidad familiar
Trabaja: 0 = No 1 = Si
Ingreso_men: Ingreso mensual
Gasto_ocio: Gasto semanal en ocio
Lugar_resid: Lugar de residencia

1 = BCN
2 = Hospitalet
3 = Altres Municipis del Barcelonès
4 = Baix Llobregat
5 = Altres Municipis
Medio_transp: Medio de transporte utilizado en el desplazamiento al centro de estudio
1 Metro
2 Autobús
3 Tren
4 Coche
5 Moto
6 Bicicleta o a pie
Tiempo_viaje: Tiempo empleado en el viaje de ida al centro de estudio
Nota_acceso: Nota de acceso a la Universidad
Asig_matric: Número de asignaturas matriculadas en el curso anterior
Asig_aprob: Número de asignaturas aprobadas del curso anterior

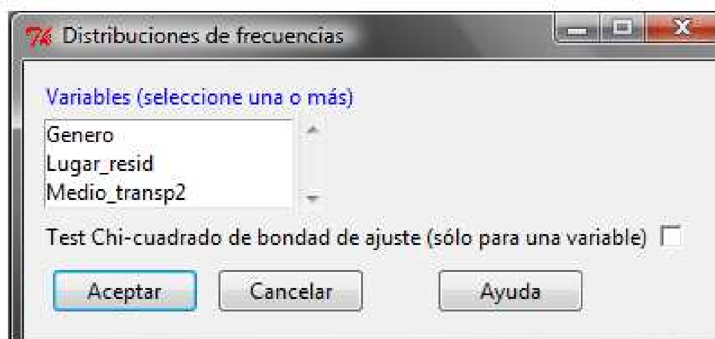
Se pide:

- Indicar la naturaleza de cada una de las características observadas.
- Realizar con la ayuda del programa R-commander la tabla de frecuencias y las representaciones gráficas de las características: Genero, Medio de Tranporte, Miembros y Nota de Acceso.
- Elaborar el diagrama de tallo y hojas de las variables: Edad, Tiempo del viaje y Nota de Acceso

Instrucciones para la realización del Actividad 2_1 con R-Commander

- Distribución de frecuencias datos cualitativos

Estadísticos ▶ Resúmenes ▶ Distribución de frecuencias



Se selecciona la variable o variables y Aceptar

- Distribución de frecuencias variable discreta (pocos valores diferentes)

Escribiendo en la ventana de instrucciones y ejecutando se obtiene:
table(Nombasedat$Nomvar), la distribución de frecuencias absolutas.
sum(table(Nombasedat$Nomvar)), el total de observaciones, sin contar los valores missing (NA)
table(Nombasedat$Nomvar)/sum(table(Nombasedat$Nomvar)), la distribución de frecuencias relativas.
100*(table(Nombasedat$Nomvar))/sum(table(Nombasedat$Nomvar)), la distribución de frecuencias relativas en porcentajes

- Distribución de frecuencias con los valores de la variable agrupados en intervalos

La instrucción puede ser:
hist(Nombasedat$Nomvar, nclass = 10, plot = FALSE)
(en nclass se debe indicar el número de intervalos que se desea). 0 también
table(cut(Nombasedat$Nomvar, breaks=c(0, 10, 20, 30, 50))

- Representación gráfica datos cualitativos

Diagrama de barras
Gráficas ► Gráfica de barras
Gráfico de sectores
Gráficas ► Gráfica de sectores
En los correspondientes cuadros se seleccionan las variables y se acepta.

- Diagrama de barras variable discreta (pocos valores diferentes)

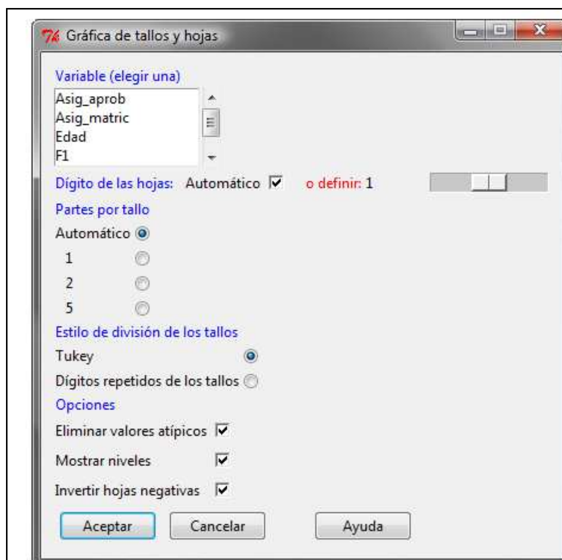
Se obtiene escribiendo en la ventana de instrucciones:
barplot(table(Nombasedat$Nomvar), xlab= "EtiquetaejeX", ylab = "EtiquetaejeY")
Los commandos **xlab** e **ylab** se incluirán cuando se quieran etiquetar los ejes de coordenadas.

- Histograma

Gráficas ► Histograma
La instrucción que se ejecuta es:
Hist(Nombasedat$Nomvar, nclass = 10, col = "grey", xlab = "EtiquetaejeX", ylab = "EtiquetaejeY")
Se puede modificar el número de intervalos, las etiquetas de los ejes, el color del gráfico. Por ejemplo, fijando el valor de col, como “grey”, “red”, “green”, etc los rectángulos del histograma estarán coloreados de gris, rojo, verde, etc.

- Diagrama de tallo y hojas (Stem and leaf plot)

Gráficas ► Gráfica de tallos y hojas



En el cuadro de diálogo:
Se selecciona la variable
Se recomienda dejar activado Automático para Dígito de las hojas y Partes por tallo.
Si a cada valor del tallo le corresponde un número muy grande de hojas, pueden repetirse los valores del tallo 2 o 5 veces, activando los números correspondientes en Partes por tallo.
Se recomienda dejar activadas las Opciones Eliminar valores atípicos y Mostrar niveles.
Una vez elegidas las características del diagrama se debe aceptar para ejecutar.

Naturaleza de cada una de las variables

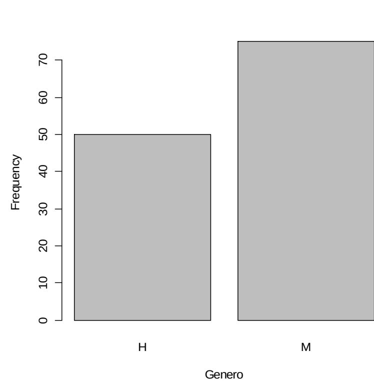
Genero	cualitativa nominal con dos modalidades
Edad	cuantitativa discreta
Miembros	cuantitativa discreta
Trabaja	cualitativa nominal con dos modalidades
Ingreso_men	cuantitativa continua
Gasto_ocio	cuantitativa continua
Lugar_resid	cualitativa nominal con varias modalidades
Medio_transp	cualitativa nominal con varias modalidades
Tiempo_viaje	cuantitativa continua
Nota_acceso	cuantitativa continua
Asig_matric	cuantitativa discreta
Asig_aprob	cuantitativa discreta

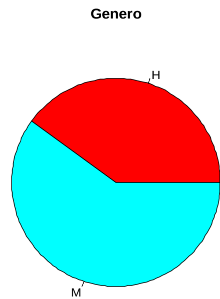
Distribución de frecuencias y representación gráfica

Estadísticos ► Resúmenes ► Distribución de frecuencias
Gráficas ► Gráfica de barras
Gráficas ► Gráfica de sectores

Genero

H	M		H	M
50	75		40	60

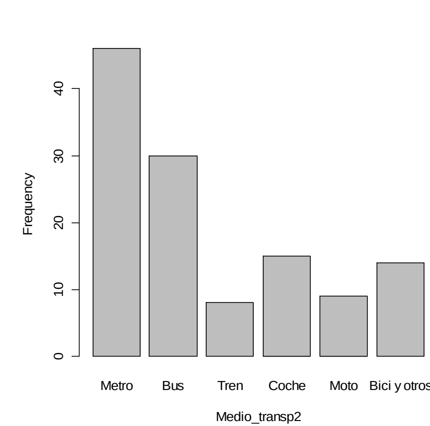


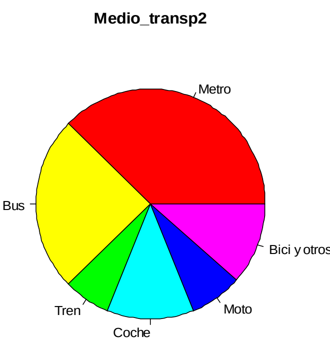


Medio de transporte

Metro	Bus	Tren	Coche	Moto	Bici y otros
46	30	8	15	9	14

Metro	Bus	Tren	Coche	Moto	Bici y otros
37.704918	24.590164	6.557377	12.295082	7.377049	11.475410





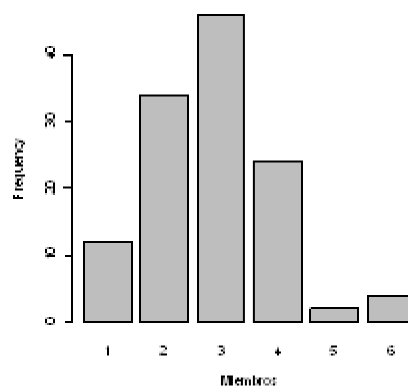
Miembros

Como Miembros es una variable numérica, discreta, con pocos valores distintos, para obtener la distribución de frecuencias y el gráfico deben introducirse y ejecutarse las instrucciones:

table(Ejercicio21$Miembros)

```
 1  2  3  4  5  6
12 34 46 24  2  4
```

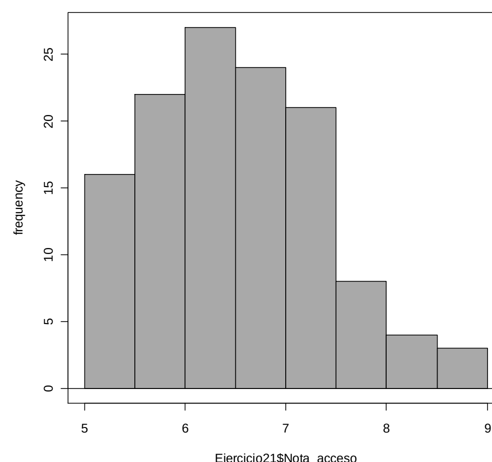
barplot(table(Ejercicio21$Miembros))



Nota de acceso

Gráficas ► Histograma

Hist(Ejercicio21$Nota_acceso, scale="frequency", breaks="Sturges", col="darkgray")



La tabla de frecuencias correspondiente al histograma anterior se obtiene ejecutando la instrucción:

hist(Ejercicio21$Nota_acceso, scale="frequency", breaks="Sturges", plot=FALSE)

El resultado es[2]:

```
$breaks
[1] 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0
$counts
[1] 16 22 27 24 21 8 4 3
$intensities
[1]   0.2559999   0.3520000   0.4320000   0.3840000   0.3360000
0.1280000 0.0640000 0.0480000
$density
[1]   0.2559999   0.3520000   0.4320000   0.3840000   0.3360000
0.1280000 0.0640000 0.0480000
$mids
[1] 5.25 5.75 6.25 6.75 7.25 7.75 8.25 8.75
```

[2] Los valores de $intensities y $density son los cocientes entre la frecuencia relativa de cada intervalo y la amplitud de éste.

3. Diagrama de tallo y hojas de las variables:

Gráficas ► Gráfica de tallo y hojas

Edad

```
1 | 2: represents 1.2
 leaf unit: 0.1
            n: 125
    1    17 | 0
   14    18 | 0000000000000
   33    19 | 0000000000000000000
   52    20 | 0000000000000000000
  (15)   21 | 000000000000000
   58    22 | 00000000000
   47    23 | 00000000000
   36    24 | 000000
   30    25 | 000000
   24    26 | 0000
   20    27 | 000
   17    28 | 00000
   12    29 | 000
    9    30 | 0
HI: 32 34 36 37 37 42 46 52
```

Tiempo_viaje

```
[1] "Warning: NA elements have been removed!!"
1 | 2: represents 12
 leaf unit: 1
            n: 122
    1    1* | 0
   12    1. | 55555555555
   25    2* | 0000000000000
   37    2. | 555555555555
   52    3* | 000000000000000
  (18)   3. | 555555555555555555
   52    4* | 00000000000000
   38    4. | 55555555555555
   24    5* | 0000000000
   14    5. | 555
   11    6* | 000000
    5    6. | 5
         7* |
    4    7. | 5
    3    8* | 0
HI: 90 90
```

Nota_acceso

```
1 | 2: represents 1.2
 leaf unit: 0.1
            n: 125
    7    5 | 0000011
   11    5 | 2333
   17    5 | 444555
   30    5 | 6666666677777
   38    5 | 88899999
   45    6 | 0000011
   53    6 | 22233333
  (20)   6 | 44444444444455555555
   52    6 | 6666666667777
   39    6 | 889
   36    7 | 0000000111111
   23    7 | 222233
   17    7 | 4455
   13    7 | 66777
    8    7 | 9
    7    8 | 00
    5    8 | 2
    4    8 | 4
    3    8 | 666
```

Actividad 2_2

Retraso en minutos de 81 vuelos Barcelona-Valencia de la compañía A

,43	1,50	2,05	2,58	2,86	2,93	3,90	4,38	4,46	5,11	5,23
5,36	5,44	5,54	6,08	6,12	6,36	6,39	6,48	6,51	6,88	7,31
7,34	8,06	8,10	8,20	8,21	8,23	8,34	8,56	8,64	8,73	8,73
8,73	8,73	8,74	8,88	8,90	8,93	9,14	9,25	9,56	9,68	9,85
9,87	9,94	9,99	10,06	10,11	10,24	10,26	10,51	11,23	11,60	11,63
11,64	11,85	11,92	12,34	12,78	12,94	13,05	13,18	13,31	13,48	13,88
14,05	14,15	14,23	14,24	14,30	14,55	14,59	15,42	15,45	15,71	16,07
16,54	16,84	17,04	19,39							

Agrupe los datos en una distribución de frecuencias y realice su representación gráfica.

1. Determinación del número de intervalos o clases.

Cuando el tamaño muestral es moderado se fija provisionalmente un número de intervalos aproximadamente igual a $\sqrt{n}$. En este caso $\sqrt{81}$ = 9, de forma que, en principio, los valores de la variable se agruparán en 9 intervalos.

2. Amplitud del intervalo.

La amplitud del intervalo, **a**, se fija aproximadamente como
a = Recorrido/ nº de clases

El recorrido de esta variable, **R**, es:
R = Valor máximo – Valor mínimo = 19,39 – 0,43 = 18,96

a = 18,96/9 = 2,10 ≈ 2

3. Límites superior e inferior de los intervalos

El valor menor de la variable es 0,43, por lo que el límite inferior del primer intervalo puede fijarse en 0. Para determinar el límite superior del primer intervalo al límite inferior se le suma la amplitud a = 2; por tanto, dicho límite será 0 + 2 = 2.

El límite inferior del segundo intervalo coincide con el límite superior del primero, es decir, 2 y su límite superior se obtiene sumándole la amplitud del intervalo a = 2, y será 4. De esta forma resultan los 9 intervalos siguientes:

0 – 2; 2 – 4, 4 – 6; 6 – 8; 8 – 10; 10 – 12; 12 – 14; 14- 16; 16 – 18

Como la variable toma valores mayores que 18 es necesario definir un intervalo adicional de límites 18 y 20; luego se agruparan los valores de la variable en 10 intervalos.

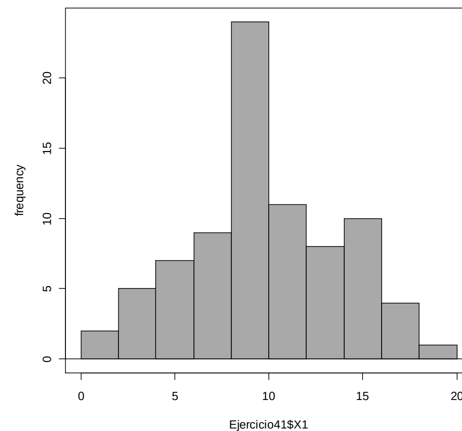
Por último, se tiene que indicar si los intervalos son abiertos o cerrados. Por omisión, se interpreta que los intervalos son $(L_{i-1}, L_i]$ cerrados por la derecha:

4. Distribución de frecuencias.

La distribución de frecuencias es la siguiente:

Retraso Xi	Frecuencia absoluta ni	Frecuencia relativa fi	Frecuencia absoluta acumulada Ni	Porcentaje acumulado
0 - 2	2	0,025	2	2,5
2 - 4	5	0,062	7	8,7
4 - 6	7	0,086	14	17,3
6 - 8	9	0,111	23	28,4
8 - 10	24	0,296	47	58,0
10 - 12	11	0,136	58	71,6
12 - 14	8	0,099	66	81,5
14 - 16	10	0,123	76	93,8
16 - 18	4	0,049	80	98,7
18 - 20	1	0,012	81	100,0
Total	81	1		

5. Representación gráfica



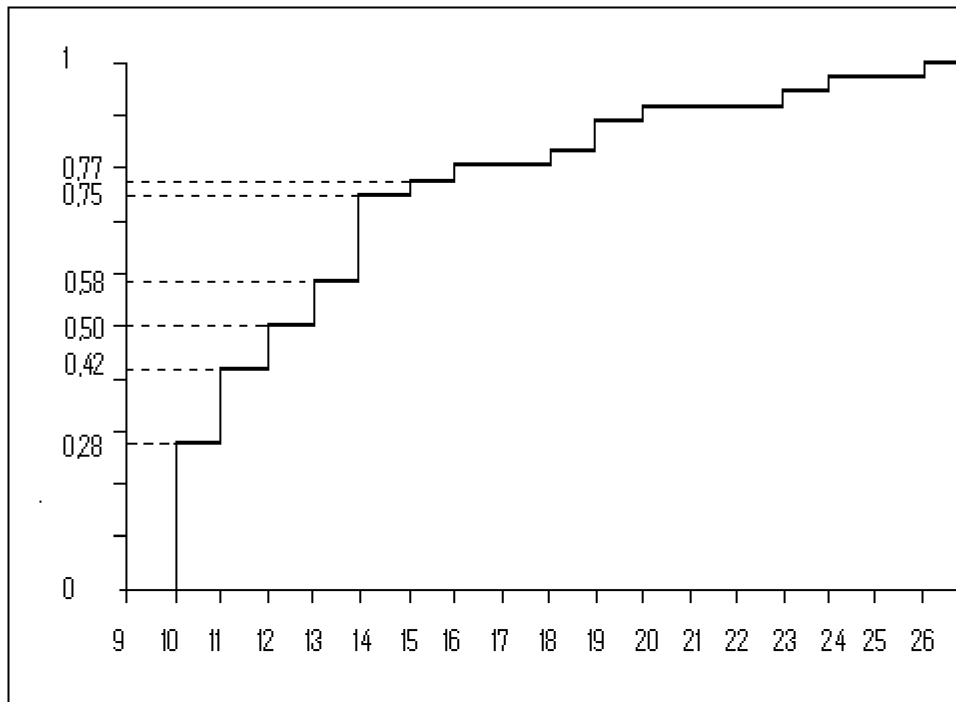
La representación gráfica correspondiente a una variable continua es el histograma

Esta distribución:

- es campaniforme (la frecuencia disminuye en los extremos)
- el centro se sitúa alrededor de 10,
- el intervalo más frecuente es (8; 10],
- no presenta lagunas,
- presenta un recorrido de aproximadamente 20 minutos.
- no hay valores anómalos u outliers

Actividad 2_3

Se entiende por punto de riesgo un cruce, un tramo o una zona, donde se han producido más de 10 accidentes. En el área metropolitana de Barcelona se han contabilizado 200 puntos de riesgo. La distribución de X= 'nº de accidentes de tráfico en estos puntos' se recoge en el siguiente diagrama de frecuencias relativas acumuladas:



1. Indique el porcentaje de puntos de riesgo que presentan:
a) exactamente 12 accidentes
b) más de 12 accidentes
c) como mínimo 14 accidentes
d) como máximo 13 accidentes

2. ¿Cuántos de estos 200 puntos de riesgo han presentado exactamente 15 accidentes?

3. ¿Podemos afirmar que el número mínimo de accidentes en el 25% de los puntos con más riesgo es 20 accidentes?

4. ¿Cuántos accidentes como máximo presenta un punto que se encuentra entre el 50% con menos riesgo?

1. Indique el porcentaje de puntos de riesgo que presentan:

a) Exactamente 12 accidentes: $f(X=12) = F(12) - F(11) = 0{,}50 - 0{,}42 = 0{,}08$ (8%)

b) Más de 12 accidentes: $f(X>12) = 1 - F(12) = 1 - 0{,}50 = 0{,}50$ (50%)

c) Como mínimo 14 accidentes: $f(X \geq 14) = 1 - F(13) = 1 - 0{,}58 = 0{,}42$ (42%)

d) Como máximo 13 accidentes: $f(X \leq 13) = F(13) = 0{,}58$ (58%)

2. ¿Cuántos de estos 200 puntos de riesgo han presentado exactamente 15 accidentes?

$f(X=15) = F(15) - F(14) = 0{,}77 - 0{,}75 = 0{,}02$

$n(X=15) = 200\ (0{,}02) = 4$

3. ¿Podemos afirmar que el número mínimo de accidentes en el 25% de los puntos con más riesgo es 20 accidentes?

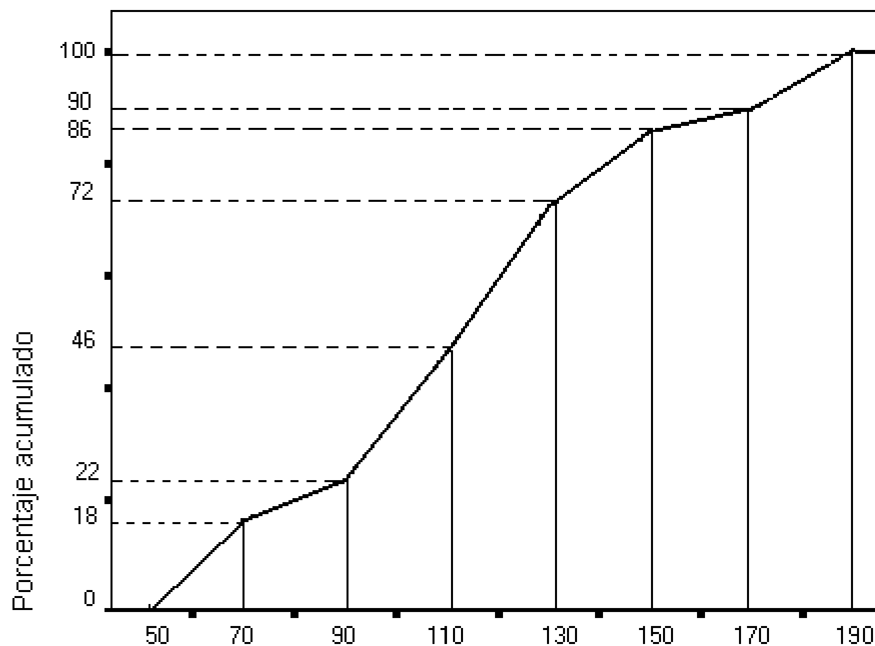
No, porque $F(20) > 0{,}75$

4. ¿Cuántos accidentes como máximo presenta un punto que se encuentra entre el 50% con menos riesgo?

Como $F(12) = 0{,}50$, el número máximo de accidentes que puede tener un punto de riesgo para estar entre el 50% de los de menos riesgo es 12 accidentes.

Actividad 2_4

Se ha observado la variable X= "Saldo (en Euros)" de 400 cuentas corrientes de clientes con edades comprendidas entre 18 y 25 años. El siguiente gráfico recoge la distribución de porcentajes acumulados de esta variable.



1. Indique el porcentaje y el número de cuentas con un saldo de:

a) Como máximo 90 Euros.
b) Más de 110 Euros.
c) Entre 90 y 130 Euros.
d) Inferior o igual a 170 Euros.

2. Una cuenta con 120 Euros ¿está entre el 60% de las que tienen más saldo?

3. Una cuenta con 100 Euros ¿está entre el 20% de las que tienen más saldo?

1. Indique el porcentaje y el número de cuentas que tienen un saldo de:

a) Como máximo 90 Euros.

$f(X \leq 90) = F(90) = 0,22$ $n(X \leq 90) = 0,22\ (400) = 88$ cuentas

b) Más de 110 Euros.

$f(X > 110) = 1 - F(110) = 1 - 0,46 = 0,54$ $n(X>110) = 0,54\ (400) = 216$ cuentas

c) Entre 90 y 130 Euros.

$f(90 \leq X \leq 130) = F(130) - F(90) = 0,72 - 0,22 = 0,50$

$n(90 \leq X \leq 130) = 0,50\ (400) = 200$ cuentas

d) Inferior o igual a 170 Euros.

$f(X \leq 170) = F(170) = 0,90$ $n(X \leq 170) = 0,90\ (400) = 360$ cuentas

2. Una cuenta con 120 Euros ¿está entre el 60% de las que tienen más saldo?

El valor mínimo del 60% con más saldo coincide con el máximo del 40% con menos saldo. Proyectando en el gráfico la frecuencia acumulada de 0,4 vemos que el valor que le corresponde es, aproximadamente, 100 €. Por lo tanto, 120 € **sí** que se encuentra entre el 60% de los cuentas con más saldo.

3. Una cuenta con 100 Euros ¿está entre el 20% de las que tienen más saldo?

El valor que corresponde a una frecuencia acumulada de 0,8 (1-0,20) es, aproximadamente, 140 €. Por lo tanto, 100 € **no** se encuentra en el 20% de los cuentas con más saldo.

Actividad 2_5

Con los siguientes diagramas Stem and Leaf conteste las preguntas.

X: Peso en gramos

```
1 | 2: represents 12
 leaf unit: 1
            n: 100
    2     110 | 58
          111 |
    4     112 | 59
    6     113 | 68
   12     114 | 024679
   16     115 | 0234
   20     116 | 8999
   28     117 | 01235699
   40     118 | 001223566689
   49     119 | 011446688
  (10)    120 | 2233556699
   41     121 | 11346667999
   30     122 | 1233456789
   20     123 | 035899
   14     124 | 01789
    9     125 | 6678
    5     126 | 12
          127 |
    3     128 | 27
    1     129 | 2
```

a) Peso mínimo X_{MIN}
b) Peso máximo X_{MAX}
c) Número de observaciones n
d) Frecuencia absoluta de125,6 gr.
e) Frecuencia absoluta acumulada hasta 1149 gr.
f) Número de paquetes en la muestra que pesan más de 1250 gr.
g) Proporción de paquetes en la muestra que pesan como máximo 1150 gr.
h) Proporción de paquetes en la muestra que pesan como mínimo 1216 gr.

Solución

a) Peso mínimo X_{MIN} 1105 gr
b) Peso máximo X_{MAX} 1292gr
c) Número de observaciones n=100 observaciones
d) Frecuencia absoluta de125,6 gr. ni = 0
e) Frecuencia absoluta acumulada hasta 1149 gr Ni = 12
f) Número de paquetes en la muestra que pesan más de 1250 gr. 9 paquetes

g) Proporción de paquetes en la muestra que pesan como máximo 1150 gr.
13%
h) Proporción de paquetes en la muestra que pesan como mínimo 1216 gr.
37%

Y: Precio en Euros

```
 leaf unit: 10
             n: 79
    8      s | 66677777
   23     0. | 888888999999999
   37     1* | 00000000001111
  (18)     t | 222222333333333333
   24      f | 444444445555555
    9      s | 6666667
HI: 350 370
```

a) Número de establecimientos observados, n
b) Precio mínimo
c) Precio máximo
d) Precio más frecuente
e) Número de establecimientos que cobran menos de 100 Euros
f) Proporción de establecimientos que cobran 90 Euros
g) Proporción de establecimientos que cobran como mínimo 140 Euros

Solución

a) Número de establecimientos observados n = 79 establecimientos
b) Precio mínimo 60 €
c) Precio máximo 370 €
d) Precio más frecuente 130 €
e) Número de establecimientos que cobran menos de 100 Euros 23 establecimientos
f) Proporción de establecimientos que cobran 90 Euros 11,39%
g) Proporción de establecimientos que cobran como mínimo 140 Euros 30,38 %

X: Cilindrada en cc

```
1 | 2: represents 1200
 leaf unit: 100
            n: 68
   9    1* | 122344444
  25    1. | 5556677777889999
  33    2* | 12222333
  (3)   2. | 599
  32    3* | 222
  29    3. | 6678
  25    4* | 000002
  19    4. | 9999
  15    5* | 222
  12    5. | 7777778
   5    6* | 3
   4    6. | 55
   2    7* | 22
```

a) Cilindrada más frecuente
b) Cilindrada máxima
c) Número de modelos de coche observados
d) Máxima cilindrada que presentan los 10 modelos menos potentes
e) Cilindrada mínima que presentan los 15 modelos más potentes

Solución

a) Cilindrada más frecuente: 5700 cc
b) Cilindrada máxima: 7200cc
c) Número de modelos de coche observados: 68 modelos
d) Máxima cilindrada que presentan los 10 modelos menos potentes 1500 cc
e) Cilindrada mínima que presentan los 15 modelos más potentes 5200 cc

EJERCICIOS TEMA 2

Ejercicio 1. La distribución del número de trabajadores en una muestra de gestorías es:

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1,00	1	1,0	1,0	1,0
	2,00	5	5,0	5,0	6,0
	3,00	12	12,0	12,0	18,0
	4,00	20	20,0	20,0	38,0
	5,00	23	23,0	23,0	61,0
	6,00	23	23,0	23,0	84,0
	7,00	12	12,0	12,0	96,0
	8,00	2	2,0	2,0	98,0
	9,00	2	2,0	2,0	100,0
	Total	100	100,0	100,0	

Determine:

a) Número de gestorías encuestadas.
b) Porcentaje de gestorías con 5 trabajadores.
c) Nº de gestorías con un máximo de 3 trabajadores.
d) Nº máximo y mínimo de trabajadores.
e) Nº más frecuente de trabajadores.
f) Nº máximo de trabajadores que tienen las 30 gestorías con menos personal.
g) Si una empresa de software únicamente está interesada en enviar propaganda a las gestorías con más de 6 empleados, ¿a qué porcentaje de las gestorías muestreadas se dirigirá?
h) Si la empresa de software está interesada en enviar propaganda al 25% de las gestorías con mayor empleo, ¿cuál es el número mínimo de empleados que debe tener una gestoría para poder estar incluida en este grupo?
i) Si el INEM se propone ayudar al 25% de las gestorías con menor empleo enviando un trabajador en prácticas, ¿cuántos empleados como máximo deberán tener para poderse beneficiar de dicha ayuda?
j) Represente gráficamente la distribución de frecuencias y compruebe las respuestas anteriores en los gráficos.

Ejercicio 2. Los siguientes datos corresponden al número de bibliotecas públicas por 1000 habitantes en las 41 comarcas de Catalunya:

25, 9, 12, 31, 57, 13, 16, 14, 22, 13, 6, 11, 12, 15, 27, 42, 9, 36, 22, 21,
7, 13, 25, 19, 16, 64, 33, 16, 43, 37, 23, 37, 19, 11, 43, 14, 49, 28, 51, 8, 9

a) Obtenga la distribución de frecuencias agrupando los datos en intervalos de amplitud 10, fijando el extremo inferior de la primera clase en 5.
b) Represente gráficamente la distribución obtenida en el apartado anterior.

Ejercicio 3. La tabla siguiente recoge la distribución de la variable X='Superficie en m^2' del conjunto de viviendas registradas en el Barcelonès con un máximo de 210 m_2:

Superficie (m^2)	Nº viviendas
(0 -30]	4 050
(30 - 60]	153 900
(60 - 90]	437 400
(90 - 120]	162 000
(120 - 150]	29 160
(150 - 180]	12 150
(180 - 210]	11 340
Total	810 000

Con esta información se pide:

a) Complete la tabla de distribución de frecuencias.
b) Represente gráficamente la distribución de frecuencias (simples y acumuladas).
c) ¿Cuántas viviendas con un máximo de 210 m2 hay censadas en el Barcelonès?
d) ¿Cuántas viviendas tienen superficies entre 60 y 90 m2?
e) ¿Cuántas viviendas tienen como máximo 60 m2?
f) ¿Cuántas viviendas superan los 150 m2?
g) ¿Qué porcentaje de viviendas tiene entre 60 y 90 m2? ¿Y como máximo 60 m2?
h) ¿Puede decirse que más del 20% de estas viviendas tienen superficies entre 90 y 120 m2?
i) ¿Cuál es la superficie máxima del 73,5% de las viviendas con menor superficie?
j) ¿Puede decirse que algo más de la cuarta parte de estas viviendas superan los 90 m2? Exactamente cuántas (en términos absolutos y en términos relativos).
k) ¿Cuál es el intervalo de superficie más frecuente?
l) Si una medida de tipo fiscal quiere abarcar al 93,5% de las viviendas con menor superficie, ¿cuál es la superficie máxima a considerar por dicha medida?
m) Si una inmobiliaria está interesada sólo en las viviendas de mayor superficie y quiere conectar con un 2,9% de las viviendas registradas, ¿cuál es la superficie mínima a considerar?
n) Una vivienda con 135 m2 ¿está, aproximadamente, entre el 5% de las registradas con mayor superficie?

Ejercicio 4. Los siguientes resultados recogen la distribución de frecuencias de la puntuación obtenida en una determinada prueba:

```
$breaks
 [1]  0  2  4  6  8  10
$counts
 [1]   5  11  19  10  5
$mids
 [1]  1  3  5  7  9
```

Realice la tabla de frecuencias simples (absolutas y relativas) y acumuladas e indique:

a) ¿Cuántos alumnos han realizado esta prueba?
b) ¿Qué porcentaje de alumnos ha obtenido como mínimo 4 puntos?
c) ¿Cuál es la puntuación mínima del 30% de las mejores puntuaciones?
d) Si Pedro ha sacado un 3, ¿puede decirse que está entre el 10% de los alumnos con peores calificaciones?

Ejercicio 5. Los siguientes datos corresponden a la observación de la variable X='número de habitantes' en 50 municipios de una Comunidad Autónoma:

2727	1418	655	4582	764	6982	2509	1527	981	2945
6218	1745	1200	982	1418	3600	4036	5345	1091	1636
982	1527	1309	3927	2727	1445	2073	3055	1636	1964
1309	2400	1636	2400	2073	4691	1200	5564	4364	1309
3382	1418	2945	2291	1745	4036	4691	873	3491	2945

a) Realice el diagrama de tallo y hojas.
b) Con el programa R-Commander obtenga el diagrama de tallo y hojas (Stem & Leaf)
c) El 30% de los municipios con menos habitantes, aproximadamente, ¿cuántos tienen como máximo?
d) Aproximadamente, ¿cuántos habitantes como mínimo tiene que tener un municipio si se encuentra en el 16% con más habitantes?

Ejercicio 6. El siguiente diagrama Stem-and-Leaf recoge el *ÍNDICE de ALFABETIZACIÓN* (IA) correspondiente a un conjunto de países en vías de desarrollo:

```
Stem-and-leaf unit = 1    1|2  represents 12
    2    1|27
    7    2|67799
   21    3|13333566667789
   (7)   4|1333358
   19    5|023444679
   10    6|0149
    6    7|4555
    2    8|12
```

a) ¿Cuál es el tamaño muestral?
b) Indique el índice de alfabetización máximo, mínimo y el más frecuente.
c) ¿Cuántos de estos países tiene un IA inferior a 48?
d) ¿En cuántos de estos países el IA no supera el valor 53?
e) ¿Cuál es el IA mínimo de los 10 países con mayor IA?

Tema 3. MEDIDAS DE POSICIÓN

Media aritmética
Moda
Mediana
Medidas de localización: cuantiles

Una vez ordenados los datos en una distribución de frecuencias se definen una serie de medidas de síntesis que permiten resumir esta información, éstas se pueden agrupar en medidas de posición (de tendencia central y de localización), de dispersión y de forma.

Las medidas de posición central sintetizan mediante un solo valor el orden de magnitud de los valores de la variable. Se definen las siguientes:

La MEDIA ARITMÉTICA es el centro de gravedad de la distribución.
La MEDIANA es el valor de la variable que corresponde al elemento central de la distribución.
La MODA es el valor más frecuente de la variable.

Las medidas de localización, CUANTILES, dividen la distribución en un cierto número de tramos con igual número de observaciones:

Los CUARTILES dividen la distribución de frecuencias en cuatro partes.
Los DECILES dividen la distribución de frecuencias en diez partes.
Los CENTILES dividen la distribución de frecuencias en cien partes.

MEDIA ARITMÉTICA, $\overline{X}$

Es la medida de tendencia central más adecuada cuando la característica observada es cuantitativa.

Se define como el cociente entre la suma de los valores de la variable observados en los elementos de la muestra y el tamaño de ésta. Si la distribución de frecuencias se presenta con los valores de la variable agrupados en intervalos, al calcular la media utilizando las correspondientes marcas de clase se obtiene un resultado aproximado.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \overline{X} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \sum_{i=1}^{k} x_i f_i$$

La media se expresa en las unidades de medida de la variable.

Propiedades:

- La media siempre toma un valor comprendido entre los valores de X mínimo y máximo observados.
- La media aritmética es el punto de equilibrio o centro de gravedad de la distribución, es decir, la suma de las desviaciones de todos los valores de la variable con respecto a la media es igual a cero.
- En el cálculo de la media se utiliza toda la información contenida en la distribución de frecuencias.
- La media de una constante es la misma constante.
- Si a todas las observaciones de la variable, X, se le aplica una transformación del tipo X'=a+bX, la media de la variable transformada $\overline{X}'$ se puede calcular en función de la media de X, siendo $\overline{X}' = a + b\overline{X}$.
- Si se divide la distribución en submuestras disjuntas y exhaustivas, la media de la distribución se puede calcular a partir de las medias de las submuestras ponderando estas últimas por el número de elementos que contienen.
  Por ejemplo, si se ha subdividido en tres grupos A, B y C, la media global es:

$$\overline{X} = \frac{\overline{X}_A n_A + \overline{X}_B n_B + \overline{X}_C n_C}{n_A + n_B + n_C}$$

Inconvenientes

- Sólo se puede obtener si la característica observada es cuantitativa.
- La media es muy sensible a la presencia de observaciones extremas, tendiendo a desplazarse hacia éstas. Cuando esto ocurre la media no sintetiza adecuadamente la distribución de la variable. En estos casos puede calcularse la MEDIA RECORTADA.

MEDIANA, Me

La mediana es el valor de la variable correspondiente al elemento que ocupa la posición central. La mediana, por tanto, divide la distribución de frecuencias en dos partes con igual número de elementos.

Características

- La mediana se expresa en las mismas unidades de medida de la variable.
- Los cambios de origen y de escala modifican la mediana.
- La mediana puede ser una medida de tendencia central más representativa que la media cuando la variable presenta valores extremos.

Inconvenientes

- Sólo se puede obtener si la característica observada es ordinal.
- En el cálculo de la mediana no se tiene en cuenta toda la información contenida en la distribución de frecuencias.

El procedimiento a seguir para localizarla depende de la forma en que se presente la ordenación de los datos: Stem and leaf, tabla de frecuencias simples, tabla de frecuencias con los valores agrupados en intervalos.

- Con la ordenación de los elementos de la muestra proporcionada por el diagrama Stem and leaf, la mediana es el valor de la variable correspondiente a la posición que deja tantas observaciones por debajo como por encima. Si el número de observaciones es n la posición central es $\frac{n+1}{2}$ y la mediana es el valor que ocupa dicha posición. Si el número de observaciones es par, la mediana se calcula como el promedio de los valores de la variable correspondiente a los dos elementos centrales.

- Con los valores dispuestos en una tabla de frecuencias simple, la mediana es el valor al que corresponde la primera frecuencia absoluta acumulada mayor o igual que $\frac{n}{2}$.
- Cuando los valores de la variable se agrupan en intervalos, el intervalo que contiene a la mediana es el primero que presenta una frecuencia absoluta acumulada igual o superior a $\frac{n}{2}$. Una vez localizado el intervalo mediano, el valor de la mediana se aproxima mediante la siguiente fórmula basada en el supuesto de que la frecuencia correspondiente a cada intervalo se distribuye uniformemente dentro de éste.

$$Me = L_{i-1} + \frac{0,5n - N_{i-1}}{n_i} a_i$$

Siendo:
L_{i-1} el límite inferior del intervalo que contiene a la mediana,
n tamaño de la muestra,
N_{i-1} la frecuencia absoluta acumulada del intervalo anterior al que contiene a la mediana,
n_i y a_i son, respectivamente, la frecuencia absoluta y la amplitud del intervalo mediano.

MODA

La moda es el valor de la variable que más veces se repite en la muestra.

Características

- La moda se expresa en las unidades de medida de la variable.
- La moda es la única medida de posición que sintetiza la distribución de frecuencias de una característica categórica nominal.

Inconvenientes

- Una distribución de frecuencias puede tener más de una moda.
- Para determinar la moda no se tiene en cuenta toda la información contenida en la distribución de frecuencias.

Para localizar la moda se busca la frecuencia (absoluta o relativa) máxima, el valor de la variable correspondiente a dicha frecuencia es la moda.

Si los valores de la variable se agrupan en intervalos, el intervalo modal es aquel al que le corresponde la frecuencia máxima. En tal caso puede tomarse la marca de clase del intervalo modal como valor aproximado de la moda.

MEDIDAS DE LOCALIZACIÓN: CUANTILES

Si se ordenan los elementos de la muestra desde el que tiene el menor valor de la variable hasta el que tiene el mayor valor, los cuantiles son los valores de la variable que dividen a la distribución en un cierto número de partes con igual número de elementos.

Los cuantiles se expresan en las mismas unidades de medida de la variable y le afectan los cambios de origen y cambios de escala.

Los cuantiles más utilizados son los cuartiles, los deciles y los centiles o percentiles.

Cuartiles

Son los tres valores de la variable, Q_1, Q_2, Q_3 que dividen la distribución en cuatro partes con igual número de observaciones.

El primer cuartil, Q_1, es el valor de la variable que deja por debajo el 25% del total de observaciones. El segundo cuartil, Q_2, es el valor de la variable que deja por debajo el 50% de las observaciones y, por tanto, coincide con la mediana. El tercer cuartil, Q_3, es el valor de la variable que deja por debajo el 75% del total de observaciones.

Entre dos cuartiles consecutivos se encuentra el 25% del total de observaciones.

El cuartil es el valor de la variable al que le corresponde la primera frecuencia absoluta acumulada mayor o igual que kn dónde k= 0,25 para Q_1; k=0,5 para Q_2 y k= 0,75 para Q_3.

Si los valores de la variable se agrupan en intervalos, el intervalo que contiene al cuartil es aquel cuya frecuencia absoluta acumulada es la primera mayor o igual que kn.

Una vez localizado el intervalo que lo contiene, el valor de Q_i se aproxima mediante la siguiente fórmula basada en el supuesto de que la frecuencia correspondiente a cada intervalo se distribuye uniformemente dentro de éste.

$$Q_i = L_{i-1} + \frac{kn - N_{i-1}}{n_i} a_i$$

Siendo k= 0,25 para Q_1; k=0,5 para Q_2 y k= 0,75 para Q_3

Deciles, Centiles o Percentiles

Los decíles son los nueve valores de la variable, D_1, D_2, …., D_8, D_9 que dividen la distribución en diez partes con igual número de observaciones.

El primer decil, D_1, es el valor de la variable que deja por debajo el 10% del total de observaciones; el segundo decil, D_2, es el valor de la variable que deja por debajo el 20% de las observaciones y así sucesivamente. El quinto decil, D_5, coincide con la mediana.

Entre dos deciles consecutivos se encuentra el 10% del total de observaciones.

Los Centiles o Percentiles son los noventa y nueve valores de la variable, C_1, C_2, …., C_{98}, C_{99} que dividen la distribución en cien partes con igual número de observaciones.

El primer centil, C_1, es el valor de la variable que deja por debajo el 1% del total de observaciones; el segundo centil, C_2, es el valor de la variable que deja por debajo el 2% de las observaciones y así sucesivamente. El quincuagésimo centil, C_{50}, coincide con la mediana.

Entre dos centiles consecutivos se encuentra el 1% del total de observaciones.

Su cálculo es análogo al de los cuartiles. Una vez localizado el intervalo, el valor aproximado del cuantil es:

$$C_i = L_{i-1} + \frac{k\,n - N_{i-1}}{n_i} a_i$$

Siendo k = 0,1; 0,2,….,0,9 para los deciles y k=0,01, 0,02, ……,0,99 para los centiles.

Actividad 3_1

Se ha observado la variable X= "Saldo (en Euros)" de 400 cuentas corrientes en una entidad bancaria correspondientes a clientes con edades comprendidas entre 18 y 25 años. La distribución de frecuencias de esta variable es la siguiente:

Saldo en €	Nºcuentas
50 – 70	72
70 - 90	16
90 – 110	96
110 – 130	104
130 – 150	56
150 – 170	16
170 - 190	40
Total	400

1. Calcule las medidas de posición central e indique la más adecuada.
2. Calcule los cuartiles
3. Indique cuál es el saldo mínimo de una cuenta para estar entre el 25% de las de mayor saldo.
4. Indique cuál es el saldo máximo de una cuenta para estar entre el 15% de las de menor saldo.
5. Indique cuál es el saldo mínimo de una cuenta para estar entre el 40% de las de mayor saldo.
6. Repita el ejercicio con la ayuda del programa R-Commander y con los datos del archivo Ejercicio31.rda

1. Medidas de posición central: media, mediana y moda.

Intervalo	x_i (c_i)	ni	N_i	$x_i n_i$
50 – 70	60	72	72	4320
70 - 90	80	16	88	1280
90 – 110	100	96	184	9600
110 – 130	120	104	288	12480
130 – 150	140	56	344	7840
150 – 170	160	16	360	2560
170 - 190	180	40	400	7200
Total		400	76	45280

Media aritmética: $\bar{X} = \frac{45280}{400} = 113,2$

Moda: I_{Mo} = (110; 130]

Mediana: I_{Me} = (110; 130] ya que $N_i \geq 200$ y $N_{i-1} < 200$

$$Me = L_{i-1} + \frac{0{,}5n - N_{i-1}}{n_i} a_i = 110 + \frac{200-184}{104} 20 = \ 113{,}08 \text{ €}$$

2. Cuartiles

Primer cuartil: I_{Q1} = (90; 110] ya que $N_i \geq 100$ y $N_{i-1} < 100$

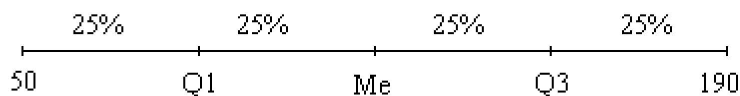
$$Q_1 = L_{i-1} + \frac{0{,}25n - N_{i-1}}{n_i} a_i = 90 + \frac{100-88}{96} 20 = 92{,}5 \text{ €}$$

Segundo cuartil: Q_2 = Me = 113,08 €

Tercer cuartil: I_{Q3} = (130; 150] ya que $N_i \geq 300$ y $N_{i-1} < 300$

$$Q_3 = L_{i-1} + \frac{0{,}75n - N_{i-1}}{n_i} a_i = 130 + \frac{300-288}{56} 20 = 134{,}28 \text{ €}$$

3. Indique cuál es el saldo mínimo que tiene que tener una cuenta para estar entre el 25% de las de mayor saldo.



Nos piden el valor $Q_{3,}$ que es 134,28 €

4. Indique cuál es el saldo máximo que tiene que tener una cuenta para estar entre el 15% de las de menor saldo.

Nos piden el valor $C_{15,}$

I_{C15} = (50; 70] ya que $N_i \geq 60$ y $N_{i-1} < 60$

$$C_{15} = L_{i-1} + \frac{0{,}15n - N_{i-1}}{n_i} a_i = 50 + \frac{60-0}{72} 20 = 66{,}67 \text{ €}$$

5. Indique cuál es el saldo mínimo que tiene que tener una cuenta para estar entre el 40% de las de mayor saldo.

Nos piden el valor D_6,

I_{D6} = (110; 130] ya que $N_i \geq 240$ y $N_{i-1} < 240$

$$D_6 = L_{i-1} + \frac{0{,}6\,n - N_{i-1}}{n_i} a_i = 110 + \frac{240-184}{104} 20 = 120{,}77\ €$$

6. R-Commander.

La secuencia para realizar el análisis descriptivo básico con R-Commander es:
Estadísticos ► Resúmenes ► Resúmenes numéricos
Para obtener el centil 15 y el decil 6 añada en cuantiles los valores 0.15 y 0.6.
La instrucción que se obtiene es:

```
numSummary(Nombase dat[,"Nomvar"], statistics=c("mean", "sd",
"quantiles"), quantiles=c(0,.25,.5,.75,1,.15,.6))
```

Los resultados obtenidos (excluyendo la desviación estándar) son:

```
mean      0%   25%  50%   75%   100%   15%    60%     n
113.595   51   93   113   136    190    67   120.4   400
```

Actividad 3_2

1. Con la base de datos **Ejercicio21.rda** y el programa R-Commander realice un análisis descriptivo (tabla de frecuencias o principales medidas de síntesis) de cada una de las siguientes variables:

Lugar de residencia
Edad
Medio de transporte
Tiempo de viaje
Notas de acceso

2. En base a los resultados obtenidos en el apartado anterior indique:

a) Lugar de residencia más frecuente.
b) Porcentaje de alumnos que utilizan los distintos medios de transportes.
c) Edad del alumno más joven.
d) Edad máxima del 25% de los alumnos más jóvenes.
e) ¿Qué porcentaje de alumnos tienen entre 19 y 21 años?
f) ¿Cree que la distribución de la variable Edad presenta valores anómalos o extremos?
g) Nota de acceso máxima.
h) Nota de acceso media.
i) Nota máxima del 50% de los alumnos peor calificados.
j) Respecto al tiempo de viaje, ¿qué porcentaje de alumnos tardan entre 25 y 45 minutos?
k) ¿Cuál es el tiempo de viaje mínimo del 25 % de los alumnos con viajes más largos?
l) En total, ¿cuál es el tiempo empleado por estos 122 alumnos?

Instrucciones para la realización del Ejercicio 3_2 con R-Commander Para las variables cualitativas, el análisis descriptivo básico consiste en la distribución de frecuencias. (Véase las instrucciones del Ejercicio 2_1) Para las variables cuantitativas, el análisis descriptivo básico se realiza con la secuencia: Estadísticos ► Resúmenes ► Resúmenes numéricos Se abre el cuadro



Se seleccionan las variables. Con las opciones activadas por defecto se obtienen, para cada variable, la media, la desviación estándar, los valores mínimo y máximo y los cuartiles.

1. Variables cualitativas:

Lugar de residencia

```
 BCN Hospitalet Altres del Barcelonès Baix Llobregat Altres municipis
  55         36                    10              8               14
```

Moda = BCN

Medio de transporte

```
Metro          Bus        Tren        Coche         Moto Bici y otros
   46           30           8           15            9           14
```

Moda = Metro

Variables cuantitativas:

```
                   mean         sd 0%   25%   50%   75%  100%   n NA
Edad          22.832000  5.5310502 17 19.00 21.00 24.00 52.00 125  0
Nota_acceso    6.496560  0.8340245  5  5.87  6.48  7.06  8.67 125  0
Tiempo_viaje 36.311475 15.4385382 10 25.00 35.00 45.00 90.00 122  3
```

2. En base a los resultados obtenidos en el apartado anterior indique:

a) Lugar de residencia más frecuente: BCN
b) Porcentaje de alumnos que utilizan los distintos medios de transporte.

```
    Metro       Bus      Tren     Coche      Moto Bici y otros
37.704918 24.590164  6.557377 12.295082  7.377049     11.475410
```

c) Edad del alumno más joven: 17 años
d) Edad máxima del 25% de los alumnos más jóvenes: 19 años
e) ¿Qué porcentaje de alumnos tienen entre 19 y 21 años?: 25%
f) ¿Cree que la distribución de la variable Edad presenta valores anómalos o extremos?
g) Si, una edad de 52 años parece un valor extremo dentro de esta distribución de edades (Véase: Media = 22,8, Mediana= 21, Q1 = 19 y Q3=24).
h) Nota de acceso máxima:8,67
i) Nota de acceso media:6,4965
j) Nota máxima del 50% de los alumnos peor calificados: 6,48
k) Respecto al tiempo de viaje, ¿qué porcentaje de alumnos tardan entre 25 y 45 minutos?: 50%
l) ¿Cuál es el tiempo de viaje mínimo del 25 % de los alumnos con viajes más largos?: 45 mn
m) En total, ¿cuál es el tiempo empleado por estos 122 alumnos?: 4430 minutos

Actividad 3_3

Determine los cuartiles de los siguientes diagramas Stem and leaf:

Y: Precio en Euros

```
1 | 2: represents 120
 leaf unit: 10
             n: 79
    8    0 | 66677777
   23    0 | 888888999999999
   37    1 | 00000000001111
  (18)   1 | 222222333333333333
   24    1 | 444444445555555
    9    1 | 6666667
HI: 350 370
```

Q_1 Posición de Q_1 = 0,25 (79) = 19,75 ≈ 20 Q_1 = 90 €

Me Posición de Me = $\frac{79+1}{2}$ = 40 Me = 120 €

Q_3 Posición de Q_3 contando desde el valor máximo 0,25 (79) = 19,75 ≈ 20 Q_3 = 140 €

X: Cilindrada en cc

```
1 | 2: represents 1200
 leaf unit: 100
             n: 68
   9    1* | 122344444
  25    1. | 5556677777889999
  33    2* | 12222333
  (3)   2. | 599
  32    3* | 222
  29    3. | 6678
  25    4* | 000002
  19    4. | 9999
  15    5* | 222
  12    5. | 7777778
   5    6* | 3
   4    6. | 55
   2    7* | 22
```

Q_1 Posición de Q_1 = 0,25 (68) = 17 Q_1 = 1700 cc

Me Posición de Me = $\frac{68+1}{2} = 34,5$ X_{34} = 2500 cc y X_{35} = 2900 cc

$$Me = \frac{2500+2900}{2} = 2700 \text{ cc}$$

Q_3 Posición de Q_3 contando desde el valor máximo 0,25 (68) =17

$Q_3 = 4900$ cc

Con el programa R Commander los resultados son:

	0%	25%	50%	75%	100%	n
Y	60	90	120	140	370	79
X	1100	1700	2700	4900	7200	68

EJERCICIOS TEMA 3

Ejercicio 1. Dada la siguiente distribución de la variable X = "Número de créditos personales concedidos en una determinada oficina bancaria" observada en una muestra de n días:

X	(0,5]	(5,10]	(10,15]	(15,20]	(20,25]	(25,30]	(30,35]
Ni	6	15	39	40	8	6	6

a) Indique el tamaño de la muestra
b) Calcule el promedio de créditos concedidos por día.
c) Calcule la media recortada de X eliminando el 10% de las observaciones extremas.

Ejercicio 2. Una empresa dedicada a la venta a domicilio ha fijado en sus 5 sucursales las siguientes dietas por vendedor:

Sucursal	Número Vendedores	Dietas(€/Vendedor)
A	11	150
B	13	140
C	19	110
D	20	150
E	17	200

En el conjunto de las 5 sucursales, ¿cuál es la dieta media por vendedor?

Ejercicio 3. Se invierten 1.000 u.m. en dos carteras compuestas por diferentes títulos de renta variable. Las cantidades invertidas y la rentabilidad de los títulos han sido las siguientes:

CARTERA A			CARTERA B		
Tít	Cant. Invertida	Rentabilidad	Tít	Cant. Invertida	Rentabilidad
A	80	4%	E	350	6%
B	170	7%	F	100	5,5%
C	130	8%	G	50	7%
D	120	5,64%			

a) ¿En cuál de las dos carteras la rentabilidad media ha sido mayor?
b) ¿Cuál es la rentabilidad media obtenida del total invertido?

Ejercicio 4. Durante el último mes se han realizado las siguientes operaciones de cambio:

A) De Libras a Euros		B) De Euros a Libras	
Tipo de Cambio (Euros/Libra)	Importe (Libras)	Tipo de Cambio (Euros/Libra)	Importe (Euros)
1,50	120	1,50	300
1,45	180	1,45	348
1,40	260	1,40	700
1,42	340	1,42	1136

Para cada uno de los casos anteriores, indique el tipo de cambio medio en Euros/Libra.

Ejercicio 5. La cotización media mensual en Bolsa de un cierto valor durante los últimos 12 meses ha sido: 70,5; 67,9; 69,1;72,3; 73,1; 74,2; 75,1; 74,8; 72,1; 71,2; 69,5 y 67,1. Si por experiencia se sabe que su comportamiento es ajustarse al valor medio, qué haría ¿comprar o vender?

Ejercicio 6. El personal de una empresa está formado por operarios y técnicos. Se sabe que 45 son operarios y suponen 3/4 partes del total de la plantilla, y el resto son técnicos. El salario medio de los operarios es 12500 u.m. y la masa salarial de los técnicos asciende a 630000 u.m. ¿Cuál es el salario medio de los trabajadores de esta empresa?

Ejercicio 7. El salario medio mensual en una determinada empresa, cuyo personal está formado únicamente por técnicos y operarios, es 1820 €. Si el salario medio mensual de los técnicos es 2100 € y el de los operarios 1700 €, ¿cuál es el porcentaje de operarios empleados?

Ejercicio 8. Se sabe que un taller produce por término medio 10 piezas/hora en el turno de día, 8 en el turno de noche y 9,2 piezas/hora considerando los dos turnos. Si el taller funciona 20 horas al día, ¿cuántas horas se trabaja en el turno de día? ¿y en el de noche?

Ejercicio 9. Un examen de una determinada asignatura consta de dos partes tales que al calcular la nota final se da doble importancia al resultado de la primera parte. Un alumno que ha obtenido un 5 en la primera parte y su nota final es un 6, ¿qué nota tenía en la segunda parte del examen?

Ejercicio 10. Calcule las medidas de tendencia central (media, mediana y moda) de las distribuciones de frecuencias de los ejercicios propuestos 1, 2 y 3 del Tema 2.

Ejercicio 11. El importe (en Euros) de 150 tickets de la cafetería de un hotel se ha tabulado y se ha obtenido:

Importe		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	(0; 50]	10	6.7	6.7	6.7
	(50; 70]	30	20.0	20.0	26.7
	(70; 90]	25	16.7	16.7	43.3
	(90; 120]	27	18.0	18.0	61.3
	(120; 150]	20	13.3	13.3	74.7
	(150; 200]	18	12.0	12.0	86.7
	(200; 250]	15	10.0	10.0	96.7
	(250; 300]	5	3.3	3.3	100.0
	Total	150	100.0	100.0	

Indique, aproximadamente:

a) Importe máximo y mínimo observado.
b) Importe máximo del 50% de los tickets con menor importe.
c) Número de tickets con un importe mínimo de 175 Euros.
d) Importe mínimo y máximo del 50% de los tickets centrales.
e) ¿Cuál es el importe mínimo de un ticket que está entre el 30% de los de mayor importe?
f) ¿Qué porcentaje de los tickets contabilizados en esta tabla tienen un importe superior a 250€?
g) Se quiere seleccionar los 9 tickets con menor importe. Aproximadamente, ¿cuál es el importe máximo de estos tickets?

Ejercicio 12. El importe (en Euros) de los últimos 58 tickets de caja de un establecimiento es:

```
1 | 2: represents 1.2
     leaf unit: 0.1
                n: 58
        2      5 | 05
        6      6 | 0589
        7      7 | 9
        9      8 | 34
       14      9 | 12345
       21     10 | 0133458
       23     11 | 33
      (16)    12 | 0122235677777899
       19     13 | 12568899
       11     14 | 01225
        6     15 | 266689
```

Conteste las siguientes cuestiones e indique el nombre del estadístico correspondiente.

a) Importe máximo del 25% de los tickets con menor importe.
b) Importe mínimo del 25% de los tickets de mayor cuantía.
c) Diferencia entre el importe mayor y el menor del 50% de los tickets centrales.
d) Importe máximo del 40% de los tickets con menor importe.
e) Se ha aplicado un descuento a los tickets con mayor importe. Si sólo se han beneficiado un 15%, ¿a partir de qué importe se ha aplicado?
f) ¿Qué porcentaje de tickets presentan un importe superior a12,8€?

Ejercicio 13. En una Cooperativa el 20% del personal es administrativo y el resto técnico. Se ha recogido información sobre la antigüedad del personal y los resultados son:

Antigüedad (en años)	Administrativos %	Técnicos %
1	10	5
2	15	20
3	25	20
4	30	40
5	20	15

a) Halle la antigüedad media del total de la plantilla.
b) En la distribución de los Administrativos, ¿cuál es la antigüedad máxima del 20% de los trabajadores con menor antigüedad?
c) En la distribución de los Técnicos, ¿cuál es la antigüedad mínima del 30% de los trabajadores con mayor antigüedad?
d) En la distribución de toda la plantilla, ¿cuál es la antigüedad máxima del 50% de los trabajadores?

Tema 4. MEDIDAS DE DISPERSIÓN Y FORMA

Recorrido. Recorrido intercuartílico.
Diagrama Box Plot (Diagrama de caja)
Varianza y desviación estándar. Coeficiente de Variación
Transformaciones lineales. Variable estandarizada
Medidas de forma

La dispersión o variabilidad de una distribución de frecuencias indica hasta que punto ésta es homogénea. Así, cuando los valores de la variable difieren poco entre sí, el grado de homogeneidad es elevado y las medidas de posición central (media) representan adecuadamente el orden de magnitud de los valores de la variable. Por el contrario, cuando entre los valores de la variable hay grandes diferencias, la distribución es heterogénea y, en consecuencia, las medidas de posición central (media) pueden ser poco representativas.

Las medidas y gráficos de dispersión que se definen en este tema son:

RANGO y RANGO INTERCUARTILICO.

DIAGRAMA BOX-PLOT. Visualiza la dispersión de una distribución en base a 5 valores: mínimo, Q1, Me, Q3 y valor máximo.

VARIANZA Y DESVIACIÓN ESTÁNDAR. Miden la dispersión de los valores de la variable respecto a la media aritmética.

COEFICIENTE DE VARIACIÓN DE PEARSON. Permite comparar el grado de dispersión relativa de varias distribuciones de frecuencias.

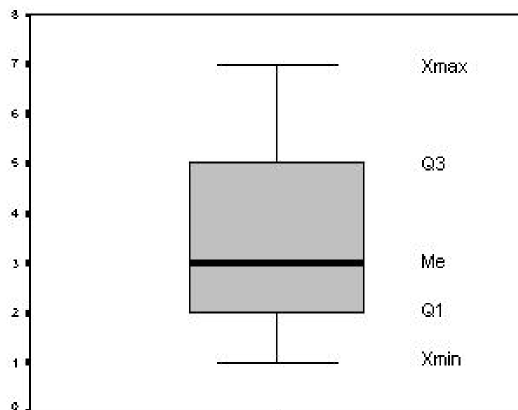
TRANSFORMACIONES LINEALES. Estudian los efectos de un cambio en el origen y/o de las unidades de medida de la variable sobre las medidas de posición y dispersión. En particular se analiza la transformación conocida por tipificación o ESTANDARIZACIÓN de una variable y sus aplicaciones.

Para completar las medidas de síntesis de las distribuciones de variables unidimensionales se interpretan medidas de la forma de la distribución: ASIMETRÍA y CURTOSIS.

DIAGRAMA BOX-PLOT (diagrama de caja)

Es un gráfico de dispersión entorno a la mediana basado en 5 valores: X_{MIN}, Q_1, Me, Q_3, X_{MAX}.

Se compone de una caja central de longitud igual al recorrido intercuartíllico y unos segmentos laterales o bigotes que abarcan el recorrido o rango de la distribución.



- Muestra la asimetría de la distribución.
- Permite comparar la dispersión y los valores centrales de varias distribuciones.
- Señala como puntos separados los valores extremos u outliers. Estos valores se clasifican en:
- ATÍPICOS, si distan del primer o tercer cuartíl en más de 1,5 veces el recorrido intercuartílico o superan los límites:
  LI = Q1-1,5 RQ y LS = Q3+1,5 RQ
- EXTREMOS, si distan del primer o tercer cuartíl en más de 3 veces el recorrido intercuartílico o superan los límites:
  LI = Q1-3 RQ y LS = Q3+3 RQ

VARIANZA, S^2

Es una medida de dispersión respecto a la media aritmética. Se define como el promedio de las desviaciones, elevadas al cuadrado, de los valores observados respecto a la media.

$$S^2=\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{n-1} \qquad S^2=\frac{\sum_{i=1}^{k}(x_i-\overline{X})^2 n_i}{n-1}$$

Se puede simplificar su cálculo utilizando la siguiente expresión:

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{X}^2}{n-1} \qquad S^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i - n\overline{X}^2}{n-1}$$

Complementa la información proporcionada por la media. Indica si $\overline{X}$ es representativa de la distribución.

Propiedades:

- La varianza siempre toma valores no negativos.
- Si todos los valores de la distribución son iguales, la varianza es 0.
- La varianza no cambia cuando se suma una misma cantidad a todos los valores observados, es decir, cuando se realiza un cambio de origen.
- La varianza se modifica si se multiplican todas las observaciones de la distribución por la misma constante, es decir, cuando se realiza un cambio de escala o cambio de unidades de medida.
- En general, si a todas las observaciones de la variable, X, se le aplica una transformación del tipo X'=a+bX, la varianza de la variable transformada $S^2_{X'}$ se puede calcular en función de la varianza de X, siendo $S^2_{X'} = b^2\, S^2_X$.

Inconvenientes:

- No presenta la misma unidad de medida que la variable.
- Depende de los cambios de la unidad de medida.
- No está acotada.

DESVIACIÓN ESTÁNDAR O TÍPICA, S

Se define como la raíz cuadrada positiva de la varianza.

$$s = +\sqrt{s^2}$$

Presenta la misma unidad de medida que la variable y que la media aritmética.

Permite establecer intervalos centrados en la media tales que como mínimo contienen un determinado porcentaje de observaciones. Por ejemplo, el siguiente intervalo alrededor de la media ($\overline{X} - 3\,s$; $\overline{X} + 3\,s$) contiene como

mínimo el 89% de las observaciones. Por lo tanto, prácticamente contiene todos los valores de la distribución y sólo algunos valores extremos superarán estos límites. Es decir, los valores de la variable sólo en algunos casos extremos diferirán de la media en más de 3 veces la desviación estándar.

La desviación estándar verifica las mismas propiedades que la varianza: es no negativa, sólo le afectan los cambios de escala y es cero cuando la distribución es constante.

COEFICIENTE DE VARIACIÓN, CV(x)

Es el cociente entre la desviación estándar y el valor absoluto de la media.

$$CV(X) = \frac{S_X}{|\bar{X}|}100$$

Expresa la desviación estándar como porcentaje de la media.

Propiedades:

- Es una medida de dispersión relativa (no presenta unidades de medida).
- Permite comparar la dispersión alrededor de la media de dos o más distribuciones aunque presenten distintas unidades de medida o medias aritméticas diferentes.
- No le afectan los cambios de unidades (cambios de escala).
- Cuando $\bar{X}=0$ el CV es indeterminado.

TRANSFORMACIONES LINEALES DE UNA VARIABLE

Transformar los valores observados de una variable cuantitativa, X, consiste en modificar cada uno de ellos mediante una misma operación aritmética obteniéndose los nuevos datos transformados, X'. Las transformaciones lineales son del tipo X'= a+bX. Contemplan:

Cambio de origen: a todas las observaciones de la variable X se les suma (resta) una constante cualquiera (a). La nueva variable será X'=X+a

Cambio de escala: todas las observaciones se multiplican (dividen) por una constante cualquiera (b). La nueva variable será X'= bX

Cambio de origen y de escala: todas las observaciones se multiplican por una constante (b) y se les suma otra constante (a). La nueva variable será X'= a+bX

Incidencia de estas transformaciones lineales en las medidas resumen:

Transformación	Media aritmética	Mediana	Moda
X'=X+a	$\bar{X}' = \bar{X} + a$	Me(X')=Me(X) +a	Mo(X')=Mo(X) +a
X'=bX	$\bar{X}' = b\bar{X}$	Me(X')= b Me(X)	Mo(X')= b Mo(X)
X'=a+bX	$\bar{X}' = a + b\bar{X}$	Me(X')= a+b Me(X)	Mo(X')= a+b Mo(X)

Transformación	Varianza	Desviación estándar	Coeficiente Variación
X'=X+a	$S^2_{X'}=S^2_X$	$S_{X'}=S_X$	$CV_{X'} \neq CV_X$
X'=bX	$S^2_{X'}=b^2 S^2_X$	$S_{X'}=\|b\| S_X$	$CV_{X'} = CV_X$
X'=a+bX	$S^2_{X'}=b^2 S^2_X$	$S_{X'}=\|b\| S_X$	$CV_{X'} \neq CV_X$

TIPIFICACIÓN O ESTANDARIZACIÓN

Es un caso particular de cambio de origen y escala. Cuando a todas las observaciones de X se les resta su media ($\bar{X}$) y la diferencia se divide por su desviación estándar (S_X) se obtiene una nueva variable Zi que recibe el nombre de variable tipificada o estandarizada.

$$Z_i = \frac{X_i - \bar{X}}{S_X}$$

El valor estandarizado, Z_i, indican el número de desviaciones estándar que el valor particular X_i está por encima (si Z es positivo) o por debajo (si Z es negativo) de la media $\bar{X}$.

Los valores estandarizados son puntuaciones adimensionales que permiten efectuar comparaciones en términos relativos de la posición de un elemento o de un valor en dos o más distribuciones.

Propiedades:

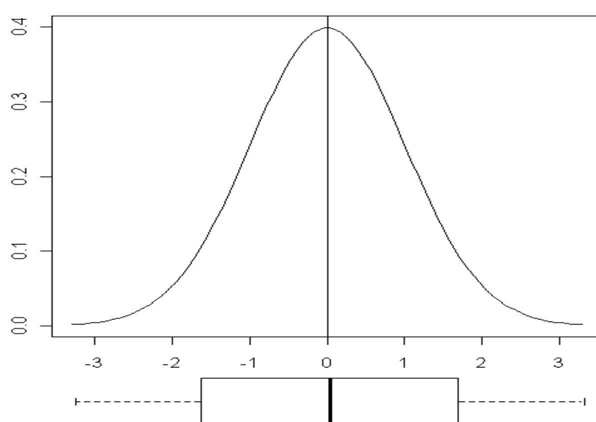
- Su media aritmética es cero: $\bar{Z} = 0$

- Su desviación estándar es uno: $S_Z=1$
- Es una variable sin unidades de medida (adimensional)

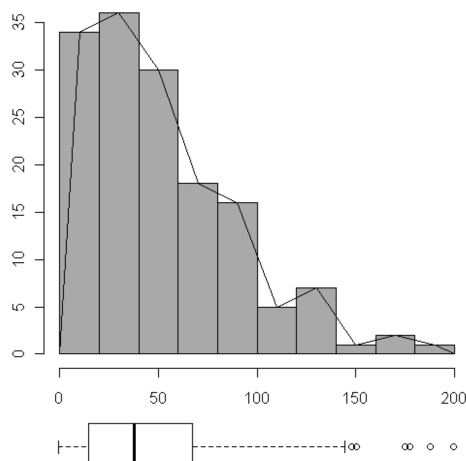
ASIMETRÍA

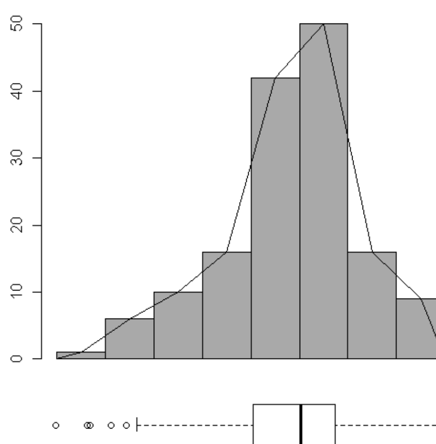
Una distribución de frecuencias es simétrica si su representación gráfica (diagrama de barras o histograma), tiene un eje de simetría perpendicular al eje de abscisas tal que la parte de la distribución que queda a un lado del eje es la imagen especular de la parte que queda al otro lado del eje. En caso contrario, la distribución es asimétrica.

La siguiente distribución es simétrica.



Las siguientes distribuciones son asimétricas:





Una distribución presenta asimetría positiva cuando la cola o los valores más alejados de la media están a la derecha. Si la cola es hacia la izquierda se

dice que tiene asimetría negativa. De las distribuciones anteriores la primera tiene asimetría positiva y la segunda asimetría negativa.

MEDIDAS DE ASIMETRÍA

Coeficiente de asimetría de Fisher

Se define como:

$$g_1 = \frac{\dfrac{\sum_{i=1}^{k}\left(X_i-\bar{X}\right)^3 n_i}{n}}{S^3}$$

- Si la distribución es simétrica g1=0
- Si hay asimetría positiva g1>0.
- Si hay asimetría negativa g1<0.

Coeficiente de asimetría de Pearson

Si la distribución tiene forma campanoide, con una sola moda y asimetría moderada, puede utilizarse como medida de asimetría el coeficiente de asimetría de Pearson:

$$A = \frac{\bar{X}-Me}{S}$$

- Si la distribución es simétrica la media es igual que la mediana y A=0.
- Si hay asimetría positiva la media es mayor que la mediana y A>0.
- Si hay asimetría negativa la media es menor que la mediana y A<0.

MEDIDAS DE CURTOSIS

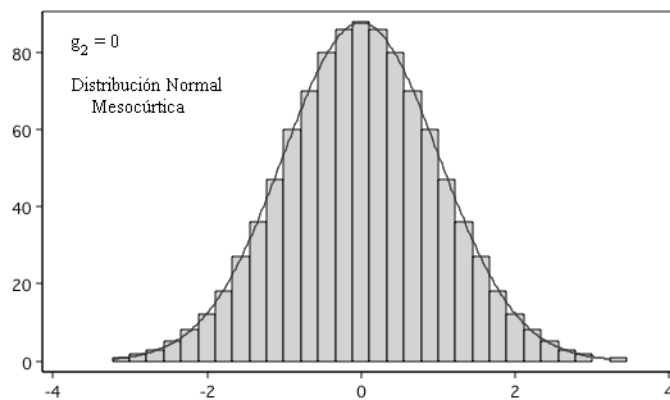
La curtosis mide el grado de apuntamiento de una distribución de frecuencias por comparación con una distribución teórica (distribución de probabilidad) de una variable continua, que recibe el nombre de distribución Normal, que se toma como modelo de referencia. Este modelo tiene forma de campana simétrica y unimodal.

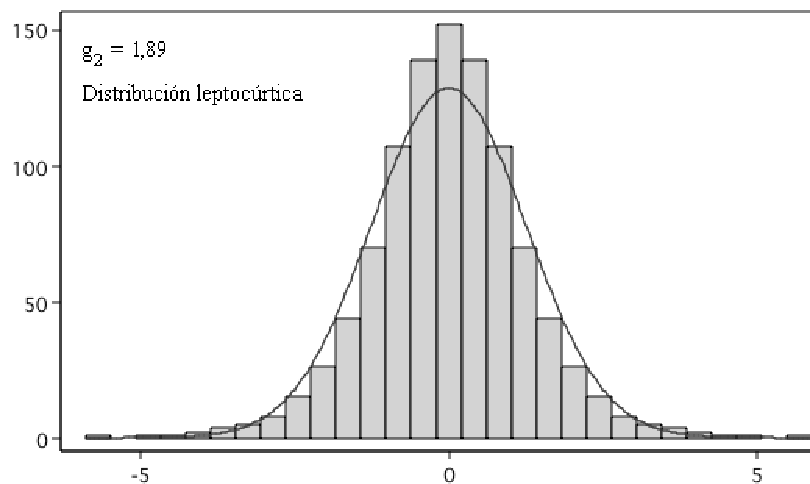
La curtosis analiza la deformación, en sentido vertical, de una distribución con respecto a la Normal.

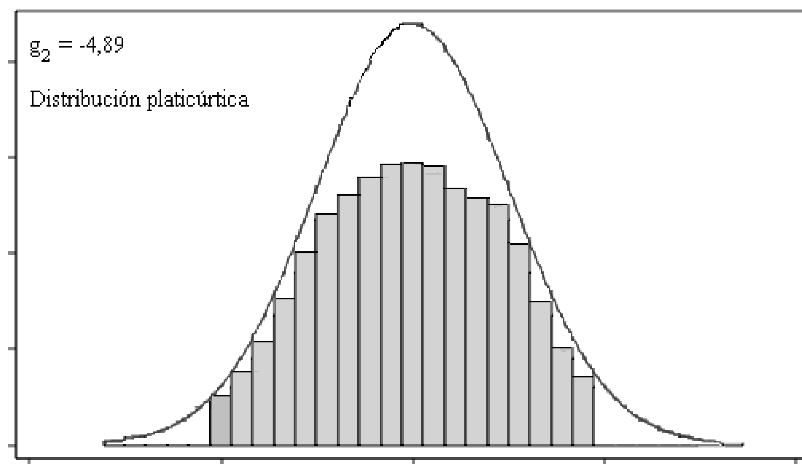
Coeficiente de Curtosis de Fisher

$$g_2 = \frac{\dfrac{\sum_{i=1}^{k}(x_i - \overline{X})^4 n_i}{n}}{S^4} - 3$$

- Si g2=0 la distribución tiene el mismo grado de apuntamiento que la curva Normal y se denomina mesocúrtica.
- Si g2>0 la distribución es leptocúrtica, y tiene mayor grado de apuntamiento que la curva Normal
- Si g2<0 la distribución es platicúrtica, y tiene menor grado de apuntamiento que la curva Normal.







La curtosis puede considerarse como una medida del peso relativo de las colas de la distribución dentro de la variación total. En una distribución leptocúrtica el peso relativo de las colas es mayor que en la distribución Normal; en una distribución platicúrtica el peso relativo de las colas es menor que en la distribución Normal.

ACTIVIDADES

Actividad 4_1

Con la tabla de frecuencias obtenida en el Actividad 2_2, realice un análisis descriptivo (medidas de tendencia central, dispersión y cuartiles) de la variable X1='Retraso en mn'.

Análisis descriptivo

Retraso (mn)	Marca de clase x_i	Frecuencia absoluta ni	Frecuencia absoluta acumulada N_i	$x_i \cdot n_i$	$x_i^2 n_i$
0 – 2	1	2	2	2	2
2 – 4	3	5	7	15	45
4 – 6	5	7	14	35	175
6 – 8	7	9	23	63	441
8 – 10	9	24	47	216	1944
10 – 12	11	11	58	121	1331
12 – 14	13	8	66	104	1352
14 – 16	15	10	76	150	2250
16 – 18	17	4	80	68	1156
18 – 20	19	1	81	19	361
Total		81		793	9057

Media y varianza

$$\overline{X}=\frac{793}{81}= 9{,}79 \text{ mn}$$

$$S^2=\frac{\sum_{i=1}^{k} x_i^2 n_i - n\overline{X}^2}{n-1}=\frac{9057-(81)9{,}79^2}{80}= 16{,}17 \qquad S = 4{,}02 \text{ mn}$$

$$CV=\frac{4{,}02}{9{,}79}= 0{,}4107$$

Cuartiles

Posiciones: $\frac{81}{4}=20{,}25$ $\qquad \frac{2(81)}{4}=40{,}5$ $\qquad \frac{3(81)}{4}=60{,}75$

La primera Ni mayor que 20,25 es 23, por tanto $Q_1 \in$ [6 – 8]

La primera Ni mayor que 40,5 es 47, por tanto Q_2 = Me ∈ [8 – 10]

La primera Ni mayor que 60,75 es 66, por tanto Q_3∈ [12 – 14]

$$Q_1=6+\frac{20,25-14}{9}2=7,39 \text{ mn} \qquad Q_2=8+\frac{40,5-23}{24}2=9,46 \text{ mn}$$

$$Q_3=12+\frac{60,75-58}{8}2=12,68 \text{ mn}$$

Dispersión

Recorrido R ≈ 20 - 0 = 20 mn

Recorrido Intercuartílico RI = 12,68 – 7,39 = 5,29 mn

Actividad 4_2

Con el programa R-Commander y la base de datos Ejercicio41.rda realice:

1. Un análisis descriptivo de la variable X1 y compare los resultados con los obtenidos en la actividad 4_1.
2. Un análisis descriptivo de X2 = 'Retraso en mn de 100 vuelos Barcelona-Valencia de otra compañía'.
3. Con los resultados obtenidos en los apartados anteriores calcule:

a) ¿Cuál es el retraso medio del total de los 181 vuelos de ambas compañías?
b) ¿En cuál de las dos distribuciones presenta la variable 'Retraso' mayor dispersión relativa?
c) Con respecto a la variable X1, determine la media, la mediana, la desviación típica y los cuartiles si el retraso se mide en segundos.
d) Compare un retraso de 8 minutos de un vuelo de la distribución X1 con un retraso de 10 minutos de un vuelo de X1 e indique cuál presenta peor posición relativa
e) Compare un retraso de 8 minutos de un vuelo de la distribución X1 con 12,43 minutos de un vuelo de X2. ¿Puede decirse que ambos vuelos ocupan la misma posición relativa dentro de sus correspondientes distribuciones?

4. Realice los diagramas Box-plot y compruebe si hay valores atípicos o extremos.

1. Análisis descriptivo de la variable X1

Instrucciones para la realización con R-Commander.
Para la obtención del análisis descriptivo básico (media, desviación estándar, cuartiles) la secuencia es:
Estadísticos ► Resúmenes ► Resúmenes numéricos.
numSummary(Ejercicio41$X1, na.rm=TRUE , statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))

```
        mean        sd    0%    25%   50%   75%  100%  n NA
X1  9.676667 4.042094 0.43  6.880  9.25 12.94 19.39  81 19
```

Para obtener la varianza, el recorrido y el recorrido intercuartílico, las instrucciones son:

Var(Ejercicio41$X1, na.rm=TRUE)
[1] 16.33852

```
range(Ejercicio41$X1, na.rm=TRUE)
[1]  0.43 19.39

IQR(Ejercicio41$X1, na.rm=TRUE)
[1] 6.06
```

Las diferencias se deben a que los valores de los estadísticos obtenidos en el apartado 1 son aproximaciones porque se han utilizado las marcas de clase en lugar de los valores observados.

2. Análisis descriptivo de la variable X2

```
        mean        sd   0%    25%   50%   75%  100%   n
X2 14.848500 5.839808 4.43 12.175 14.25 17.48 55.00 100
Var(Ejercicio41$X2)
[1] 34.10335

range(Ejercicio41$X2)
[1]  4.43 55.00

IQR(Ejercicio41$X2)
[1] 5.305
```

3. Calcule:

a) ¿Cuál es el retraso medio del total de los 181 vuelos de ambas compañías?

$$\bar{X}_{Total} = \frac{81(9,6767)+100(14,8485)}{81+100} = 12,534 \text{ mn}$$

b) ¿En cuál de las dos distribuciones presenta la variable 'Retraso' mayor dispersión relativa?

$$CV1 = \frac{4,04209}{9,6767} = 0,4177 \qquad CV2 = \frac{5,8398}{14,8485} = 0,3932$$

La variable X1 presenta mayor dispersión relativa que X2, porque CV1 > CV2

c) Con respecto a la variable X1, determinar la media, la mediana, la desviación típica y los cuartiles si el retraso se mide en segundos.

Sea Y1 = 'Retraso de los vuelos en segundos'. Y1 = 60(X1)

$\overline{Y}_1 = 60\ (9{,}6767) = 580{,}602$ sg

$Me_{Y_1} = 60\ (9{,}25) = 555$ sg

$S_{Y_1} = 60\ (4{,}04209) = 242{,}5254$ sg

Q1 = 60 (6,88) = 412,8 sg

Q3 = 60 (12,94) = 776,4 sg

d) Compare un retraso de 8 minutos de un vuelo de la distribución X1 con un retraso de 10 minutos de un vuelo de X2 e indique cuál presenta peor posición relativa

Los correspondientes valores estandarizados (tipificados) son:

$$z_1 = \frac{8 - 9{,}6767}{4{,}04209} = -0{,}4148 \qquad z_2 = \frac{10 - 14{,}8485}{5{,}8398} = -0{,}8302$$

Un vuelo con 8 minutos de retraso en X1 presenta peor posición dentro de su distribución que un vuelo de X2 con un retraso de 10 minutos, porque z1 > z2

e) Compare un retraso de 8 minutos de un vuelo de la distribución X1 con 12,43 minutos de un vuelo de X2. ¿Puede decirse que ambos vuelos ocupan la misma posición relativa dentro de sus correspondientes distribuciones?

Estandarizando ambos valores se tiene:

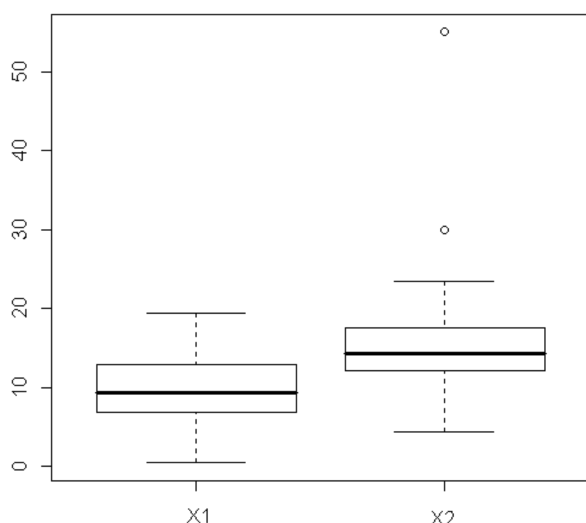
$$z_1 = -0{,}4148 \qquad z_2 = \frac{12{,}43 - 14{,}8485}{5{,}8398} = -0{,}4141$$

Los valores estandarizados son prácticamente iguales; puede decirse que los dos vuelos ocupan la misma posición relativa dentro de su correspondiente distribución.

4. Diagramas Box-plot

Instrucciones para la realización del diagrama Box-plot con R-Commander la secuencia es:

Gráficas ▶ Diagrama de caja
En el cuadro se selecciona la variable; si se activa Identificar atípicos con el ratón, haciendo clic con el botón izquierdo sobre los círculos de los valores atípicos aparecerá el número del elemento al que corresponde cada valor atípico.
Si se quiere obtener los dos diagramas de caja en el mismo gráfico, la instrucción es:
boxplot(Ejercicio41$X1, Ejercicio41$X2)



En la distribución de X2 hay dos valores 'outliers' o anómalos.
En la distribución de X2 el recorrido intercuartílico es:
RI=Q3−Q1=5,305

- Los límites que determinan los valores atípicos son:

LI=Q1−1,5RI=4,21 y LS=Q3+1,5RI=25,44
Por lo tanto, valores inferiores a 4,21 o superiores a 25,44 son atípicos. Los valores 30 y 55 sobrepasan el límite LS.

- Los límites que determinan los valores extremos son:

LI=Q1−3RI = -3,74 y LS=Q3+3RI = 33,39
Por lo tanto, valores inferiores a -3,74 o superiores a 33,39 son extremos. El valor 55 sobrepasa el límite LS.
En consecuencia, el valor 30 mn es valor atípico y 55 mn es valor extremo.

Actividad 4_3

La base de datos Ejercicio42.rda contiene observaciones de 3 variables: X_1, X_2 y X_3.

X_1 = 'Número de caramelos de limón en cada una de las 100 bolsas de 10 caramelos de la marca A'
X_2 = 'Número de caramelos de limón en cada una de las 100 bolsas de 10 caramelos de la marca B'
X_3 = 'Número de caramelos de limón en cada una de las 100 bolsas de 10 caramelos de la marca C'

Describa analítica y gráficamente la forma de las distribuciones de estas variables.

Instrucciones R-Commander.
Para calcular los coeficientes de asimetría, g1, y de curtosis, g2, es necesario activar sus correspondientes opciones junto al tipo 3 en el menú que aparece con la secuencia:
Estadísticos ► Resúmenes ► Resúmenes numéricos.

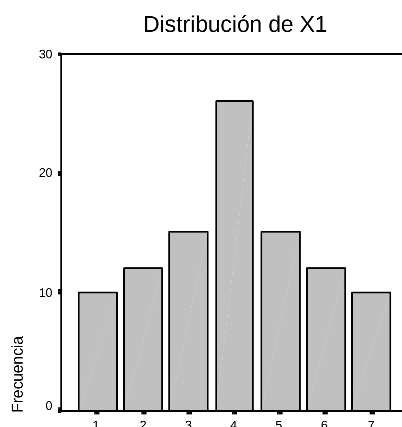
```
numSummary(Ejercicio41$X1, na.rm=TRUE, statistics=c("mean", "sd", "quantiles", "skewness",
"kurtosis"), quantiles=c(0,.25,.5,.75,1), type="3")
```

Análisis descriptivo de X1

X_1 = 'Número de caramelos de limón en cada una de las 100 bolsas de 10 caramelos de la marca A' tiene la siguiente distribución de frecuencias:

X_1	1	2	3	4	5	6	7	Total
Frecuencia	10	12	15	26	15	12	10	n = 100

Representación gráfica



La distribución de frecuencias de X_1 es simétrica.

Estadísticos descriptivos

```
mean   sd   skewness   kurtosis   0% 25% 50% 75% 100%  n
 4   1.7580    0      -0.8709862  1   3   4   5   7   100
```

La distribución de X_1 es simétrica y el coeficiente de asimetría es igual a 0. Como se trata de una distribución campanoide puede utilizarse también el coeficiente de asimetría de Pearson:

$$As = \frac{\overline{X} - Me}{S} = \frac{4-4}{1{,}758} = 0$$

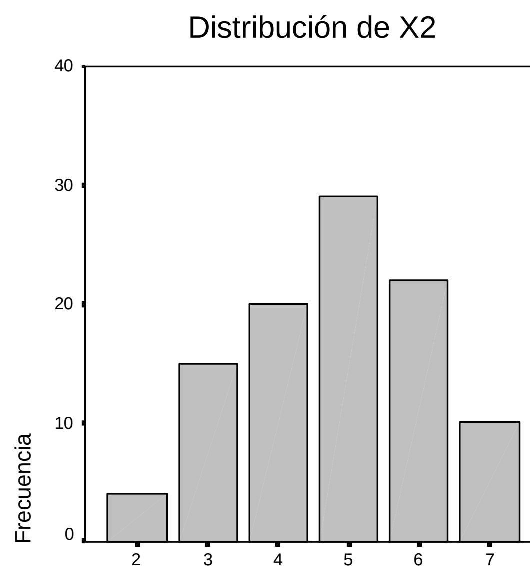
En este tipo de distribuciones $\overline{X} = Me$

Análisis descriptivo de x2

X_2 = 'Número de caramelos de limón en cada una de las 100 bolsas de 10 caramelos de la marca B' tiene la siguiente distribución de frecuencias:

X_2	2	3	4	5	6	7	Total
Frecuencia	4	15	20	29	22	10	n = 100

Representación gráfica



La distribución de X_2 presenta asimetría hacia la izquierda.

Estadísticos descriptivos

```
mean    sd      skewness    kurtosis   0% 25% 50% 75% 100%  n
4.8  1.325736 -0.1751012 -0.7616583 2  4   5   6   7  100
```

El coeficiente de asimetría es negativo, como corresponde a una distribución con asimetría hacia la izquierda o negativa. Como se trata de una distribución campanoide puede utilizarse el coeficiente de asimetría de Pearson:

$$As=\frac{\overline{X}-Me}{S}=\frac{4,8-5}{1,326}=-0,150$$

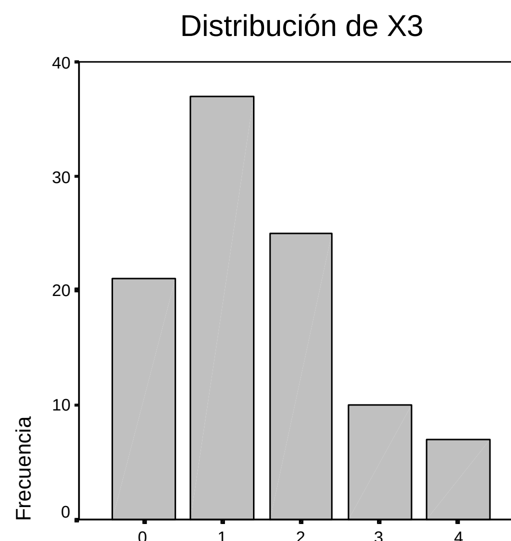
En este tipo de distribuciones $\overline{X} < Me$

Análisis descriptivo de X3

X_3 = 'Número de caramelos de limón en cada una de las 100 bolsas de 10 caramelos de la marca C' tiene la siguiente distribución de frecuencias:

X_3	0	1	2	3	4	Total
Frecuencia	21	37	25	10	7	n = 100

Representación gráfica



La distribución de X3 presenta asimetría hacia la derecha.

Estadísticos descriptivos

```
mean     sd    skewness  kurtosis   0% 25% 50% 75% 100%   n
1.45 1.140397 0.607346 -0.3373762   0   1   1   2    4  100
```

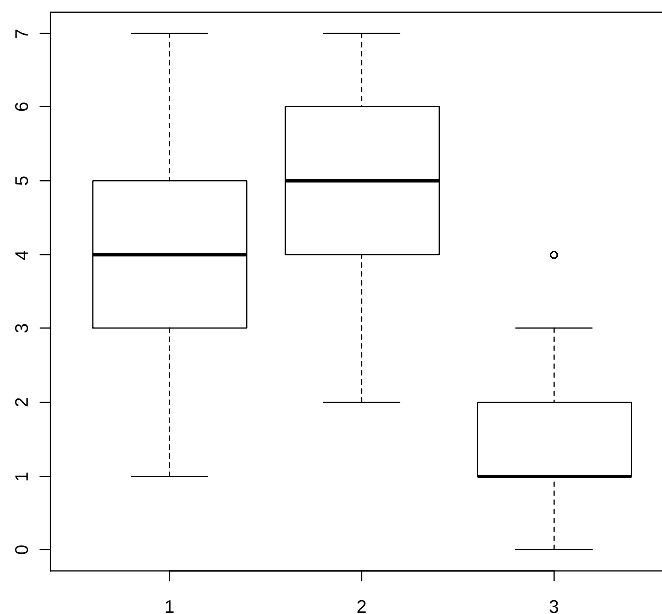
El coeficiente de asimetría es positivo, como corresponde a una distribución con asimetría hacia la derecha o positiva. Como se trata de una distribución campanoide puede utilizarse el coeficiente de asimetría de Pearson:

$$As=\frac{\overline{X}\text{-Me}}{S}=\frac{1{,}45-1}{1{,}140}=0{,}394$$

En este tipo de distribuciones $\overline{X}$ > Me

Box-plot

Con la instrucción: boxplot(Ejercicio42$X1, Ejercicio42$X2, Ejercicio42$X3)



EJERCICIOS TEMA 4

Ejercicio 1. Calcule las medidas de dispersión (absoluta y relativa) de las distribuciones de frecuencias de los ejercicios propuestos 1 y 2 del Tema 2 y ejercicio propuesto 1 del Tema 3. Interprete estos resultados.

Ejercicio 2. Dadas las siguientes distribuciones de frecuencias, calcule:

Número semanal de visitas en los servicios de urgencias de la Comunidad A (en miles)

		Frecuencia	Porcentaje
Válidos	(20-40]	30	18.8
	(40-60]	33	20.6
	(60-80]	32	20.0
	(80-100]	29	18.1
	(100-120]	21	13.1
	(120-140]	5	3.1
	(140-160]	10	6.3
	Total	160	100.0

Número semanal de visitas en los servicios de urgencias de la Comunidad B (en miles)

		Frecuencia	Porcentaje
Válidos	(50-70]	25	15.6
	(70-90]	12	7.5
	(90-110]	37	23.1
	(110-130]	42	26.3
	(130-150]	21	13.1
	(150-170]	7	4.4
	(170-190]	16	10.0
	Total	160	100.0

m) Número medio de visitas semanales en cada comunidad y en el conjunto de ambas comunidades.
n) Desviación estándar del número de visitas semanales en cada una de estas comunidades.
o) Indique en cuál de las dos distribuciones es más representativa la media.

Ejercicio 3. Los resultados de la prueba de acceso a la Universidad correspondientes a los alumnos de cuatro institutos de Barcelona (A, B, C y D) se resumen en la siguiente tabla:

	mean	sd	0%	25%	50%	75%	100%	n
A	6.62	1.85	4.01	5.28	6.97	8.47	10.00	130
B	6.48	1.68	4.31	5.28	6.13	8.71	10.00	50
C	6.05	1.46	4.02	5.60	6.93	8.22	9.97	110
D	5.99	1.95	4.01	6.65	8.22	9.32	10.00	140

Se pide:

a) Nota media del total de alumnos.
b) Indique en cuál de los cuatro institutos las calificaciones de los alumnos son más homogéneas.
c) ¿Qué transformación se debería realizar en las calificaciones de los alumnos del instituto D para que tengan la misma nota media que los del instituto A y varianza igual a 1.

Ejercicio 4. El salario medio semanal de los empleados de una empresa es 550 € con desviación típica 300 €. La empresa tiene 500 trabajadores y se plantea realizar un incremento salarial para lo cual propone las siguientes alternativas:

A. Efectuar un aumento lineal de 80 €.
B. Incrementar un 6%
C. Un incremento del 4% más un aumento lineal de 60 €.
D. Aumentar 50 € más un 5% sobre el salario que resulte después de aumentar los 50 €.

Se pide:

a) Salarios medios tras aplicar las alternativas anteriores.
b) ¿Cuál es el aumento en € de la masa salarial resultante en cada alternativa?
c) Indique cuál de las cuatro propuestas de revisión salarial disminuye la dispersión relativa del salario.

Ejercicio 5. Una empresa de telefonía controla el número de clientes captados en sus cuatro agencias. Con la información recogida en cada agencia se han obtenido los siguientes resultados de la variable X="Número de clientes captados por empleado":

	mean	sd	0%	25%	50%	75%	100%
Agencia 1	5.106	1.81	1	3	5	6	10
Agencia 2	6.712	2.63	0	5	6	8	16
Agencia 3	7.575	2.89	1	5	8	11	15
Agencia 4	8.787	3.02	4	8	9	12	16

a) Si el número de clientes captados por el último empleado contratado en cada una de estas agencias (1, 2, 3 y 4) ha sido 6, 7, 8 y 10, respectivamente, ¿cuál de estos empleados ocupa mejor posición relativa en su respectiva agencia?
b) Un empleado de la agencia 2 en la escala estandarizada del "número de clientes captados" presenta una puntuación de -1,38. ¿Cuántos clientes captó dicho empleado?
c) Dibuje el box plot de cada agencia e indique que distribución:
   - Es más dispersa
   - Es más simétrica
   - Presenta valores outliers
   - Presenta más curtosis.

Ejercicio 6. Una cadena de televisión presenta un nivel medio de audiencia (en miles de espectadores) de 1.450 en la programación de tarde y de 2.350 en la programación de noche. Si en el conjunto de cadenas de televisión la media y la varianza de la audiencia en las franjas horarias anteriores son,

respectivamente, 1.850 y 160.000 en la primera, y 2.600 y 40.000 en la segunda, ¿en qué franja ocupa esta cadena mejor posición relativa?

Ejercicio 7. En el año 2009 las bibliotecas de Cataluña disponían de un promedio por comarca de 2896 volúmenes por 1000 habitantes con una desviación típica de 1407. Sabiendo que las comarcas Garraf, Barcelonès, Tarragonès, Baix Llobregat y Val d'Aran tenían en la escala estandarizada puntuaciones de: −0,77, 0,78, 0,36, −1,3 y −1, respectivamente

a) Indique en qué comarca/s el número de volúmenes por 1000 habitantes estaba por encima de la media de Catalunya.
b) Indique si alguna de estas comarcas presenta un número de volúmenes por 1000 habitantes outlier o anómalo.
c) Calcule el número de volúmenes por 1000 habitantes de estas comarcas

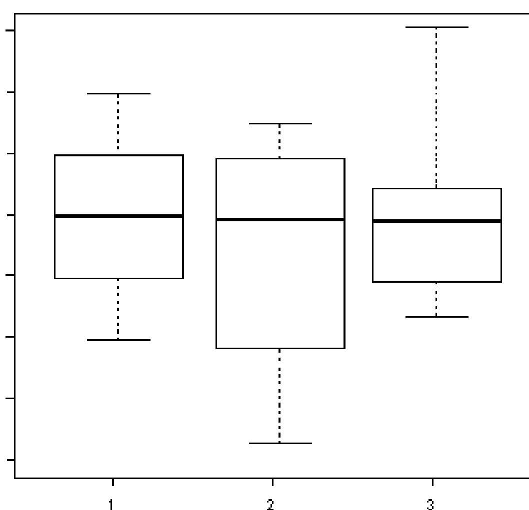
Ejercicio 8. El siguiente diagrama Stem-and-Leaf recoge el *ÍNDICE de ALFABETIZACIÓN* (IA) correspondiente a un conjunto de países en vías de desarrollo:

```
1 | 2: represents 12
 leaf unit: 1
           n: 44
   3    2* | 234
   7    2. | 5566
   8    3* | 3
   9    3. | 5
  10    4* | 3
  15    4. | 56889
  (6)   5* | 112223
  23    5. | 556677788899
  11    6* | 1122
   7    6. | 5677789
```

a) Obtenga el diagrama Box-Plot.
b) Comente las características más importantes de esta distribución.

Ejercicio 9. A partir de los resultados siguientes indique el Box-plot que corresponde a cada una de las siguientes variables:

Variable	X1	X2	X3
Mediana	20	20	20
Q_3-Q_1	16,6	10,2	6,3
Coef. Asimetría	-2,856	0	3,29
Coef. Curtosis	-0,873	1,165	2,77



Ejercicio 10. De una compañía aérea A se sabe que por término medio el 20% de los vuelos no presentan retraso, mientras que los vuelos con retraso presentan la siguiente distribución de frecuencias:

Retraso (mn)	Núm. de Vuelos
5	1000
10	1000
15	500
20	300
30	200

Se pide:

a) ¿Cuál es el retraso medio de los vuelos con retraso? ¿y del total de vuelos de la compañía?
b) Si el último vuelo realizado se encuentra entre el 20% de los vuelos con más retraso de la distribución de los vuelos totales, ¿cuál es el mínimo retraso que puede haber presentado?
c) Si otra compañía aérea B presenta los siguientes resultados:

Retraso medio del total de vuelos: 10 mn
Varianza del total de vuelos: 60

Indique en qué compañía un vuelo con un retraso de 15 mn, presentará peor posición relativa.

d) ¿En qué compañía, A o B, los retrasos presentan más dispersión relativa?
e) Si tras una política de incentivos en la Cia. A el retraso de los vuelos ha experimentado una reducción del 5%, ¿cuál será la nueva media aritmética

de los vuelos con retraso y su desviación estándar? y ¿cuál será la media del total de vuelos?

Ejercicio 11. La base de datos **Pasajeros.rda** contiene observaciones de la variable X= "Número de pasajeros en n vuelos de una compañía aérea". Explique el resultado que se obtiene ejecutando cada una de las siguientes instrucciones con el programa R-Commander (cargue previamente la base de datos **Pasajeros.rda).**

a) M=cut(Pasajeros$x,
   breaks=c(100,150,200,250,300,350,400,450,500,550,600,650))
   table(M)
b) table(M)/length(Pasajeros$x)
c) Estadísticos ► Resúmenes ► Resúmenes numéricos.
   Seleccionar: Media, Desviación típica, Asimetría, Apuntamiento, Cuantiles
   numSummary(Pasajeros[,"x"], statistics=c("mean", "sd", "quantiles", "skewness", kurtosis"), quantiles=c(0,.25,.5,.75,1), type="2")
d) Varianza: var(Pasajeros$x)
e) Recorrido intercuartílico: IQR(Pasajeros$x)
f) Media recortada eliminando el 20% de las observaciones extremas: mean(Pasajeros$x, 0.2)
g) Realice los siguientes gráficos con las opciones del menú *Gráficas*: Histograma; Gráfica de tallos y hojas; Diagrama de Caja

```
M=cut(Pasajeros$x,
breaks=c(100,150,200,250,300,350,400,450,500,550,600,650))
table(M)
```

Distribución de frecuencias absolutas con los valores de X agrupados en intervalos

```
(100,150] (150,200] (200,250] (250,300] (300,350] (350,400] (400,450] (450,500] (500,550]
       24        24        21        13        21        13        13         8         4
(550,600] (600,650]
        1         2
```

```
table(M)/length(Pasajeros$x)
```

Distribución de frecuencias relativas

```
  (100,150]   (150,200]   (200,250]   (250,300]   (300,350]   (350,400]   (400,450]
0.166666667 0.166666667 0.145833333 0.090277778 0.145833333 0.090277778 0.090277778
  (450,500]   (500,550]   (550,600]   (600,650]
0.055555556 0.027777778 0.006944444 0.013888889
```

numSummary(Pasajeros[,"x"], statistics=c("mean", "sd", "quantiles", "skewness", kurtosis"), quantiles=c(0,.25,.5,.75,1), type="2")

La media, la desviación estándar, El coeficiente de asimetría g1, El coeficiente de curtosis g2 , los valores mínimo y máximo, los cuartiles y el tamaño de la muestra son:

```
    mean       sd  skewness   kurtosis  0%  25%   50%   75%  100%   n
280.2986 119.9663 0.5831605 -0.3649419 104  180 265.5 360.5   622 144
```

var(Pasajeros$x)

```
[1] 14391.92
```

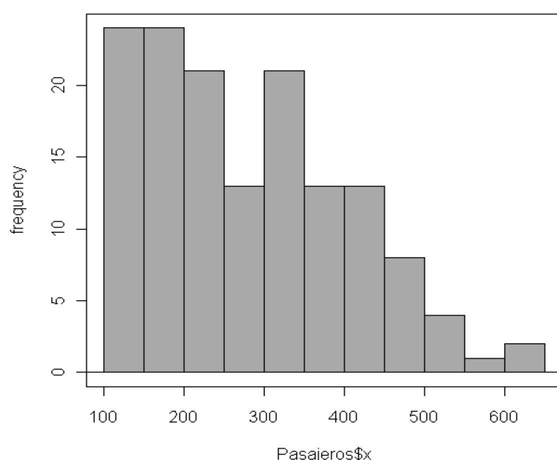
IQR(Pasajeros$x)

```
[1] 180.5
```

mean(Pasajeros$x, 0.2)

```
[1] 267.625
```

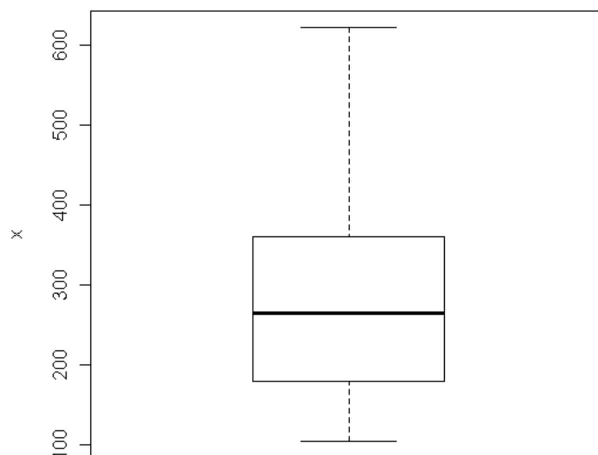
Gráficas: Histograma



Gráficas: Gráfica de tallos y hoja

```
> stem.leaf(Pasajeros$x, na.rm=TRUE)
1 | 2: represents 120
 leaf unit: 10
            n: 144
   23    1* | 01111112222333334444444
   48    1. | 5566677777788888899999999
   69    2* | 000011222233333333444
  (13)   2. | 5666677777789
   62    3* | 000001111111133444444
   41    3. | 5555666667999
   28    4* | 0000001112233
   15    4. | 66666779
    7    5* | 0034
    3    5. | 5
    2    6* | 02
```

Gráficas: Diagrama de Caja



Tema 5. DISTRIBUCIÓN DE FRECUENCIAS BIDIMENSIONALES

Distribución de frecuencias conjuntas
Distribuciones marginales
Distribuciones condicionadas
Independencia estadística

La DISTRIBUCIÓN DE FRECUENCIAS CONJUNTA proporciona tres tipos de información; en primer lugar información acerca del patrón de comportamiento conjunto de ambas variables; en segundo lugar, información por separado de cada una de las dos variables observadas (DISTRIBUCIONES MARGINALES); y, en tercer lugar, información acerca del comportamiento de una de las variables cuando se controla el valor de la otra (DISTRIBUCIONES CONDICIONADAS).

A partir del análisis de las distribuciones marginales y de las distribuciones condicionadas se define el concepto de INDEPENDENCIA ESTADÍSTICA.

DISTRIBUCIÓN DE FRECUENCIAS CONJUNTAS

Cuando sobre los n elementos de la muestra se observan simultáneamente dos variables, X e Y, cada elemento está representado por el par ordenado (x_i, y_j), donde x_i es el valor que toma la variable X e y_j es el valor que toma la variable Y.

ELEMENTOS DE LA TABLA DE DOBLE ENTRADA

X \ Y	y_1	y_2	...	y_j	...	y_J	$n(x_i)$
x_1	n_{11} (f_{11})	n_{12} (f_{12})	...	n_{1j} (f_{1j})	...	n_{1J} (f_{1J})	$n(x_1)$ $f(x_1)$
x_2	n_{21} (f_{21})	n_{22} (f_{22})	...	n_{2j} (f_{2j})	...	n_{2J} (f_{2J})	$n(x_2)$ $f(x_2)$
...	...	...	...	...	...	...	...
x_i	n_{i1} (f_{i1})	n_{i2} (f_{i2})	...	n_{ij} (f_{ij})	...	n_{iJ} (f_{iJ})	$n(x_i)$ $f(x_i)$
...	...	...	...	...	...	...	...
x_I	n_{I1} (f_{I1})	n_{I2} (f_{I2})	...	...	...	n_{IJ} (f_{IJ})	$n(x_I)$ $f(x_I)$
n(y) f(y)	$n(y_1)$ $f(y_1)$	$n(y_2)$ $f(y_2)$	...	$n(y_j)$ $f(y_j)$	...	$n(y_J)$ $f(y_J)$	n 1

Interpretación de la tabla:

x_i / y_j: en el margen izquierdo de la tabla se recogen los I valores de la variable X. En el margen superior de la tabla se recogen los J valores de la variable Y

Si X o Y o ambas son variables continuas, en los márgenes izquierdo y superior de la tabla se recogerán los intervalos en los que se agrupan los valores de cada variable.

n_{ij}: frecuencias absolutas conjuntas; siendo n_{ij} el número de elementos de la muestra para los que X = x_i e Y = y_j. La suma de todas las frecuencias absolutas conjuntas es igual a n:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = n$$

f_{ij}: frecuencias relativas conjuntas; siendo f_{ij} = n_{ij}/n la proporción en tanto por uno de elementos para los que x = x_i e Y = y_j. La suma de todas las frecuencias relativas conjuntas es igual a 1:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} = 1$$

Si las frecuencias relativas conjuntas se multiplican por 100 se obtienen los correspondientes porcentajes.

$n(x_i)$: frecuencia marginal de X. La suma de las frecuencias absolutas (relativas) conjuntas de la fila i-ésima es igual a la frecuencia absoluta (relativa) correspondiente al valor x_i:

$$\sum_{j=1}^{J} n_{ij} = n(x_i) \qquad \sum_{j=1}^{J} f_{ij} = f(x_i)$$

$n(y_i)$: frecuencia marginal de Y. La suma de las frecuencias absolutas (relativas) conjuntas de la columna j-ésima es igual a la frecuencia absoluta (relativa) correspondiente al valor y_j:

$$\sum_{i=1}^{I} n_{ij} = n(y_i) \qquad \sum_{j=1}^{J} f_{ij} = f(y_i)$$

DISTRIBUCIONES DE FRECUENCIAS MARGINALES

A partir de la distribución de frecuencias conjuntas se pueden establecer las distribuciones de frecuencias unidimensionales de X y de Y, que se llaman "marginales" porque los valores de las variables y las correspondientes frecuencias absolutas o relativas se encuentran en los márgenes de la tabla.

Distribuciones de frecuencias marginales de X y de Y

X	n(X)	f(X)
x_1	$n(x_1)$	$f(x_1)$
x_2	$n(x_2)$	$f(x_2)$
...	...	...
x_i	$n(x_i)$	$f(x_i)$
...	...	...
X_I	$n(x_I)$	$f(x_I)$
Total	n	1

Y	n(Y)	f(Y)
y_1	$n(y_1)$	$f(y_1)$
y_2	$n(y_2)$	$f(y_2)$
...	...	...
y_j	$n(y_j)$	$f(y_j)$
...	...	...
Y_J	$n(y_J)$	$f(y_J)$
Total	n	1

DISTRIBUCIONES DE FRECUENCIAS CONDICIONADAS

A partir de la distribución de frecuencias conjuntas puede establecerse el comportamiento de una de las variables, por ejemplo X, cuando la otra, Y, cumple determinada condición. Por ejemplo, la distribución de X condicionada a $Y = y_j$ es:

$X/Y = y_j$	Frec. absolutas	Frec. relativas
x_1	n_{1j}	$f_{1/j} = n_{1j}/n(y_j)$
x_2	n_{2j}	$f_{2/j} = n_{2j}/n(y_j)$
...	...	...
x_i	n_{ij}	$f_{j/j} = n_{ij}/n(y_j)$
...	...	...
x_I	n_{Ij}	$f_{I/j} = n_{Ij}/n(y_j)$
Total	$n(y_j)$	1

Análogamente, la distribución de Y condicionada a $X = x_i$ es:

Y/X = xi	Frec. absolutas	Frec. relativas
y_1	n_{i1}	$f_{1/i}= n_{i1}/n(x_i)$
y_2	n_{i2}	$f_{2/i} = n_{i2}/n(x_i)$
...	...	...
y_j	n_{ij}	$f_{j/i} = n_{ij}/n(x_i)$
...	...	...
y_J	n_{iJ}	$f_{J/i} = n_{iJ}/n(x_i)$
Total	$n(x_i)$	1

INDEPENDENCIA ESTADÍSTICA

Dos variables X e Y son estadísticamente independientes si y sólo si cada frecuencia relativa conjunta es igual al producto de las correspondientes frecuencias relativas marginales:

$$f_{ij} = f\left(x_i\right).f\left(y_j\right) \qquad \forall i,j$$

Teniendo en cuenta la relación entre frecuencias absolutas y relativas, la condición de independencia puede expresarse como:

$$n_{i,j} = \frac{n(x_i)n(y_j)}{n} \qquad \forall i,j \text{ si X e Y son independientes:}$$

Todas las distribuciones de frecuencias relativas de X condicionada a cualquier valor de Y son iguales a la distribución de frecuencias relativas marginal de X.

Todas las distribuciones de frecuencias relativas de Y condicionada a cualquier valor de X son iguales a la distribución de frecuencias relativas marginal de Y.

ACTIVIDADES

Actividad 5_1

Con la base de datos **Ejercicio51.rda** y el programa R-Commander obtenga la siguiente tabla de contingencia:

```
                Tiempo_viaje_2
Medio_transp2   0-20 20-40 40-60 60-80 80-100
  Metro           11    23    12     0      0
  Bus              5    15     8     2      0
  Tren             0     3     5     0      0
  Coche            0    11     1     1      2
  Moto             6     1     2     0      0
  Bici y otros     3     6     5     0      0
```

Indique:

f) ¿Cuántos entrevistados utilizan el transporte público y qué porcentaje representan sobre el total?
g) ¿Cuántos van en metro y tardan menos de 20 minutos?
h) ¿Qué porcentaje tarda más de 40 minutos y va en coche?
i) De los que van en autobús, ¿cuántos tardan menos de 20 minutos? y ¿qué porcentaje representan en este colectivo?
j) De los que tardan más de 20 minutos, ¿qué porcentaje va en transporte público?
k) Con respecto a los que tardan como máximo 20 minutos, ¿qué porcentaje utilizan el transporte público?
l) ¿Qué porcentaje tarda entre 20 y 40 minutos y utilizan coche o moto?
m) ¿Qué porcentaje utiliza el tren?
n) ¿Cuál es el tiempo mediano?
o) ¿Cuál es el medio de transporte más utilizado?
p) Comparando el colectivo que utiliza el coche con el que utiliza la moto, ¿en cuál de ellos es superior el tiempo máximo del 50% de las personas que tardan menos?
q) Los que hacen recorridos largos (más de 40 minutos) ¿qué medio de transporte utilizan con más frecuencia?
r) Los que utilizan el metro, en promedio ¿cuánto tiempo emplean?

La distribución de frecuencias absolutas conjuntas se obtiene con la secuencia:

Estadísticos ►*Tablas de contingencia* ► *Tabla de doble entrada*, eligiendo en el cuadro de diálogo en *Variable de fila* Medio_transp y en *Variable de columna* Tiempo_viaje y con la opción *Sin porcentaje,* que está activada por defecto.

Medio_transp \ Tiempo_viaje	0-20	20-40	40-60	60-80	80-100
Metro	11	23	12	0	0
Bus	5	15	8	2	0
Tren	0	3	5	0	0
Coche	0	11	1	1	2
Moto	6	1	2	0	0
Bici y otros	3	6	5	0	0

Si se quiere la distribución de frecuencias relativas conjuntas en porcentajes se debe activar la opción *Porcentajes totales*.

	0-20	20-40	40-60	60-80	80-100	Total
Metro	9.0	18.9	9.8	0.0	0.0	37.7
Bus	4.1	12.3	6.6	1.6	0.0	24.6
Tren	0.0	2.5	4.1	0.0	0.0	6.6
Coche	0.0	9.0	0.8	0.8	1.6	12.3
Moto	4.9	0.8	1.6	0.0	0.0	7.4
Bici y otros	2.5	4.9	4.1	0.0	0.0	11.5
Total	20.5	48.4	27.0	2.5	1.6	100.0

Si se quieren las distribuciones de X condicionadas a Y debe activarse la opción *Porcentajes por columnas*.

Medio_transp \ Tiempo_viaje	0-20	20-40	40-60	60-80	80-100
Metro	44	39.0	36.4	0.0	0
Bus	20	25.4	24.2	66.7	0
Tren	0	5.1	15.2	0.0	0
Coche	0	18.6	3.0	33.3	100
Moto	24	1.7	6.1	0.0	0
Bici y otros	12	10.2	15.2	0.0	0
Total	100	100.0	100.1	100.0	100
Count	25	59.0	33.0	3.0	2

Si se quieren las distribuciones de Y condicionadas a X debe activarse la opción *Porcentajes por filas*.

Medio_transp \ Tiempo_viaje	0-20	20-40	40-60	60-80	80-100	Total	Count
Metro	23.9	50.0	26.1	0.0	0.0	100.0	46
Bus	16.7	50.0	26.7	6.7	0.0	100.1	30
Tren	0.0	37.5	62.5	0.0	0.0	100.0	8
Coche	0.0	73.3	6.7	6.7	13.3	100.0	15
Moto	66.7	11.1	22.2	0.0	0.0	100.0	9
Bici y otros	21.4	42.9	35.7	0.0	0.0	100.0	14

Para obtener las distribuciones marginales se deben ejecutar las siguientes instrucciones:

```
T <- xtabs(~Medio_transp+Tiempo_viaje, data=Ejercicio51)
margin.table(T,1)
```

margin.table(T,2)
siendo T el nombre que le asignamos a la tabla.

Los resultados son:

```
Medio_transp
       Metro    Bus  Tren   Coche   Moto Bici y otros
          46     30     8      15      9           14


Tiempo_viaje
  0-20  20-40  40-60  60-80 80-100
    25     59     33      3      2
```

El tamaño muestral, n, se obtiene con la instrucción:
sum(T)

```
[1] 122
```

a) ¿Cuántos entrevistados utilizan el transporte público y qué porcentaje representan sobre el total?

n(Metro) + n(Bus) + n(Tren) = 84 personas
f(Metro) + f(Bus) + f(Tren) = 84/122 = 0,688 68,85%

b) ¿Cuántos van en metro y tardan menos de 20 minutos?

n(Metro,0-20) = 11 personas

c) ¿Qué porcentaje tarda más de 40 minutos y va en coche?

f(Coche, 40-100) = 0,8 + 0,8 + 1,6 = 3,2%

d) De los que van en autobús, ¿cuántos tardan menos de 20 minutos? y ¿qué porcentaje representan en este colectivo?

n(Bus, 0-20) = 5 personas
n(Bus, 0-20)/n(Bus) = 5/30 = 0,167 16,67%

e) De los que tardan más de 20 minutos, ¿qué porcentaje va en transporte público?

n(Público, 20-100)/n(20-100) = (23+12+15+8+2+3+5)/97 = 0,701
70,1%

f) Con respecto a los que tardan como máximo 20 minutos, ¿qué porcentaje utilizan el transporte público?

n(Público, 0-20)/n(0-20) = (11+5)/25 = 0,64   64%

g) ¿Qué porcentaje tarda entre 20 y 40 minutos y utilizan coche o moto?

f(Coche, 20-40) + f(Moto, 20-40) = 9,0+0,8 = 9,8 %

h) ¿Qué porcentaje utiliza el tren?

6,6%

i) ¿Cuál es el tiempo mediano?

I_{Me}= (20-40]   Me = $20+\frac{122/2 - 25}{59}20$ = 32,2 mn

j) ¿Cuál es el medio de transporte más utilizado?

M_o = Metro

k) Comparando el colectivo que utiliza el coche con el que utiliza la moto, ¿en cuál de ellos es superior el tiempo máximo del 50% de las personas que tardan menos?

En la distribución condicionada que utiliza el coche I_{Me} = (20-40]
En la distribución condicionada que utiliza la moto I_{Me} = (0-20]
Es mayor en Coche.

l) Los que hacen recorridos largos (más de 40 minutos) ¿qué medio de transporte utilizan con más frecuencia?

I (40-100] = Metro

m) Los que utilizan el metro, en promedio, ¿cuánto tiempo emplean?

$\bar{X}_{/metro} = \frac{10 \cdot 11+30 \cdot 23+50 \cdot 12}{46} =$ 30,43 mn

Actividad 5_2

Sobre una muestra de 120 chaquetas se observan las variables X = 'Duración del control de calidad (en mn)' e Y = 'Nº de defectos' y se obtiene la siguiente distribución de frecuencias conjuntas:

Tabla de contingencia X * Y

			Y			
			0	1	2	Total
X	5	Recuento	2	6	15	23
		% del total	1,7%	5,0%	12,5%	19,2%
	6	Recuento	5	10	12	27
		% del total	4,2%	8,3%	10,0%	22,5%
	7	Recuento	10	28	6	44
		% del total	8,3%	23,3%	5,0%	36,7%
	8	Recuento	12	8	6	26
		% del total	10,0%	6,7%	5,0%	21,7%
Total		Recuento	29	52	39	120
		% del total	24,2%	43,3%	32,5%	100,0%

a) Obtenga las distribuciones marginales de X e Y.
b) Obtenga las distribuciones condicionadas de Y respecto a cada uno de los valores de X.
c) Razone si existe algún tipo de asociación entre X e Y a la vista de las distribuciones condicionadas del apartado anterior.

1. Distribuciones marginales de X e Y

Valores de X	Frecuencia	Porcentaje	Porcentaje acumulado
5	23	19,2	19,2
6	27	22,5	41,7
7	44	36,7	78,3
8	26	21,7	100,0
Total	120	100,0	

N	Válidos	120
	Perdidos	0
Media		6,61
Mediana		7,00
Moda		7
Desv. típ.		1,031
Varianza		1,064
Percentiles	25	6,00
	50	7,00
	75	7,00

Valores de Y	Frecuencia	Porcentaje	Porcentaje acumulado
0	29	24,2	24,2
1	52	43,3	67,5
2	39	32,5	100,0
Total	120	100,0	

N	Válidos	120
	Perdidos	0
Media		1,08
Mediana		1,00
Moda		1
Desv. típ.		,751
Varianza		,564
Percentiles	25	1,00
	50	1,00
	75	2,00

2. Distribuciones condicionadas de Y con respecto a X

Y/X=5		Frecuencia	Porcentaje
Válidos	0	2	8,7
	1	6	26,1
	2	15	65,2
	Total	23	100,0

N	Válidos	23
	Perdidos	0
Media		1,57
Desv. típ.		,662
Varianza		,439

Y/X=6		Frecuencia	Porcentaje
Válidos	0	5	18,5
	1	10	37,0
	2	12	44,4
	Total	27	100,0

N	Válidos	27
	Perdidos	0
Media		1,26
Desv. típ.		,764
Varianza		,584

Y/X=7		Frecuencia	Porcentaje
Válidos	0	10	22,7
	1	28	63,6
	2	6	13,6
	Total	44	100,0

N	Válidos	44
	Perdidos	0
Media		,91
Desv. típ.		,603
Varianza		,364

Y/X=8		Frecuencia	Porcentaje
Válidos	0	12	46,2
	1	8	30,8
	2	6	23,1
	Total	26	100,0

N	Válidos	26
	Perdidos	0
Media		,77
Desv. típ.		,815
Varianza		,665

3. Razone si existe algún tipo de asociación entre X e Y a la vista de las distribuciones condicionadas del apartado anterior.

Se observa que:

Las distribuciones de frecuencias relativas de Y condicionada a los distintos valores de X son diferentes entre si y distintas de la distribución marginal de Y, luego X e Y no son independientes.

A medida que aumenta el valor de X = 'Duración del control de calidad' disminuye la media de Y = 'Nº de defectos', lo que indica la existencia de un patrón de comportamiento conjunto de las dos variables.

EJERCICIOS TEMA 5

Ejercicio 1. Los siguientes datos corresponden a una muestra de 30 personas en paro siendo X=Edad e Y=Género:

Edad	Género	Edad	Género	Edad	Género	Edad	Género	Edad	Género
18	H	26	H	34	M	45	M	51	H
19	M	27	H	36	H	47	M	52	H
21	H	29	M	37	M	48	M	54	H
22	M	30	H	39	H	48	H	57	M
24	H	32	M	41	H	50	M	58	H
25	H	33	H	44	M	50	H	60	H

Tabule estos datos agrupando los valores de la variable Edad en las siguientes categorías:

Edad = De 16 a 19, de 20 a 24, de 25 a 54 y más de 54

Ejercicio 2. La distribución de frecuencias relativas conjunta de las variables X= 'Actividad realizada durante el tiempo libre en un fin de semana' e Y= 'Edad' observada en un colectivo formado por 6000 personas es:

Activitats realitzades	[15-30)	[30-45)	[45-65)	≥65	Total
passejar	2.6	4.4	5.5	7.6	20.1
mirar TV o vídeo	2.9	3.2	4.0	4.2	14.3
reunions família/amics	3.7	3.1	2.9	3.7	13.4
platja/piscina	3.4	3.2	2.7	2.2	11.5
llegir	1.9	3.0	2.7	2.3	9.9
feines llar/cuinar	0.8	2.1	2.9	3.7	9.5
dormir/reposar	1.6	2.1	2.2	2.0	7.9
fer esport	2.3	1.7	1.4	0.9	6.3
sortida a bars	2.5	0.6	0.4	0.2	3.7
estudiar/classes	2.7	0.4	0.2	0.1	3.4
Total	24.4	23.8	24.9	26.9	100

Indique:

a) ¿Qué porcentaje del colectivo observado son mayores de 65 años y la actividad realizada ha sido "leer"?
b) ¿Qué porcentaje del colectivo de mayores de 65 años ha dedicado su tiempo libre a "leer"?
c) ¿Qué porcentaje del colectivo observado tiene una edad inferior a 30 años y la actividad realizada ha sido "deporte"?

d) ¿Qué porcentaje del colectivo que ha realizado deporte tiene menos de 30 años?
e) ¿Qué porcentaje del colectivo observado tiene menos de 45 años y la actividad realizada ha sido "pasear" o "leer"?
f) Del colectivo de edad inferior a 45 ¿qué porcentaje ha "paseado" o "leido"?
g) ¿Cuál es la actividad más frecuente?
h) ¿Qué grupo de edad predomina en el colectivo observado?
i) ¿Cuál es la edad mediana?
j) Entre los que tienen edad inferior a 45, ¿cuál es la actividad menos preferida?
k) Las personas con menos de 30 años, ¿representan un mayor porcentaje en el colectivo de "playa-piscina" o en el colectivo de "deportes"?
l) Hacer deporte o pasear, ¿es más frecuente en el colectivo de edad inferior a 30 años o en el de edad superior a 65?

Ejercicio 3. La siguiente tabla de contingencia recoge información sobre la estructura de 100 familias elegidas al azar en cinco países de la UE.

	España	Francia	Italia	Portugal	Grecia
UP	16	33	25	17	22
ASN	57	44	52	53	54
UAN	9	4	3	4	6
DAN	18	19	20	26	18

UP: una persona; ASN: adultos sin niños; UAN: un adulto y niños; DAN: dos adultos y niños.

a) Analice si hay independencia entre estas dos categorías.
b) ¿Cuál es la proporción de familias UAN en el conjunto formado por España y Portugal?
c) ¿Qué porcentaje representan las familias ASN portuguesas sobre la muestra?
d) Indique cuales serían las frecuencias conjuntas (teóricas) bajo el supuesto de independencia y manteniéndose las frecuencias marginales.

Ejercicio 4. Se sabe que las variables M = "Máquina utilizada" y X = "Número de piezas defectuosas producidas en un día" son estadísticamente independientes.
Se ha observado un total de 400 días y se han obtenido las siguientes distribuciones marginales:

M	M1	M2	M3	M4
f(M)	0.15	0.15	0.30	0.40

X	0	1	2	3	4
f(X)	0.25	0.30	0.25	0.15	0.05

Indique:

a) Proporción de días en los que se han observado 3 piezas defectuosas.
b) Proporción de días en que se ha observado la máquina 2.
c) Número de días en los que se han observado 0 piezas defectuosas.
d) Número de días en que se ha observado la máquina 1.
e) Número de días en que se ha observado la máquina 4 y 3 piezas defectuosas.
f) Número de días que en la máquina 1 se han observado 2 piezas defectuosas.
g) Número de piezas defectuosas observado con mayor frecuencia en la máquina 3.
h) Número máximo de piezas defectuosas que presenta el 50% de los días con menor número de defectuosas observados en la máquina 3.
i) Proporción de días en que se ha observado la máquina 2 y 0 piezas defectuosas.
j) Del total de días en que se ha observado la máquina 2, ¿qué proporción ha habido con 0 piezas defectuosas?
k) Del total de días con 2 piezas defectuosas, ¿qué proporción corresponden a la máquina 3?
l) ¿En qué máquina se han observado 4 piezas defectuosas en un mayor número de días?

Ejercicio 5. Para analizar la aceptación de dos nuevos modelos de motocicleta se ha observado durante los 25 días laborables del último mes las unidades vendidas en un concesionario:

X= Unidades vendidas del modelo A

Y= Unidades vendidas del modelo B

X (Modelo A)	Y (Modelo B)	Núm. días
0	3	1
1	1	5
2	1	10
3	2	9

Se pide:

a) Tabla de doble entrada.
b) Total de unidades vendidas de cada modelo en el período observado.

Tema 6. ASOCIACIÓN ENTRE VARIABLES

Diagrama de dispersión
Asociación lineal. Covarianza
Coeficiente de correlación de Pearson
Regresión lineal

El análisis de algunos fenómenos precisa de la observación simultánea de dos características en un determinado colectivo con el objetivo de determinar si existe algún tipo de relación o asociación entre ellas.

En aquellas situaciones en que las variables son cuantitativas el tipo de asociación se puede analizar:

Gráficamente con el DIAGRAMA DE DISPERSIÓN o nube de puntos. Éste permite visualizar la existencia o no de un patrón de comportamiento conjunto entre dos variables.

Cuantitativamente con la COVARIANZA y el COEFICIENTE DE CORRELACIÓN DE PEARSON.

La covarianza cuantifica e indica si la relación es de tipo lineal, directa o inversa; sin embargo presenta el inconveniente de que queda afectada por los cambios de escala.

El coeficiente de correlación lineal**,** como medida adimensional y acotada, mide la intensidad o el grado de asociación lineal.

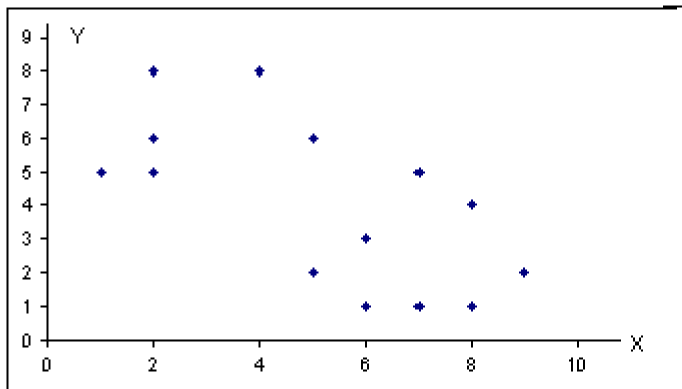
Las medidas síntesis bidimensionales de la información recogida pueden concretarse en el vector de medias y en la matriz de varianzas y covarianzas.

Por último, la RECTA DE REGRESIÓN es la recta que mejor se ajusta a la nube de puntos. El análisis de regresión implica una decisión acerca de la relación de causalidad existente entre las variables, de forma que al efectuar el ajuste es preciso decidir previamente cuál de las dos variables es la VARIABLE INDEPENDIENTE (X), es decir, cuál de ellas condiciona el comportamiento de la otra que se tomará como VARIABLE DEPENDIENTE (Y). El COEFICIENTE DE DETERMINACIÓN mide la bondad del ajuste que permite determinar si la recta ajustada explica adecuadamente la relación de dependencia existente entre X e Y.

MEDIDAS DE ASOCIACIÓN LINEAL

Análisis Gráfico

Para analizar si existe alguna relación entre las variables (X, Y) es aconsejable realizar, en primer lugar, la representación gráfica o DIAGRAMA DE DISPERSIÓN o Nube de Puntos. Se construye representando cada elemento (xi, yi) por un punto en el plano de manera que sus coordenadas son los valores que toman las dos variables.



Análisis Cuantitativo

Las principales medidas de asociación lineal para datos cuantitativos son: COVARIANZA y COFICIENTE DE CORRELACIÓN DE PEARSON

COVARIANZA, S_{XY}

Indica si existe asociación lineal y su signo.

Si se calcula con distribución de frecuencias conjuntas no unitarias:

$$S_{XY} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h}(X_i - \overline{X}).(Y_j - \overline{Y})n_{ij}}{n-1}$$

Si se calcula con distribución de frecuencias conjuntas unitarias:

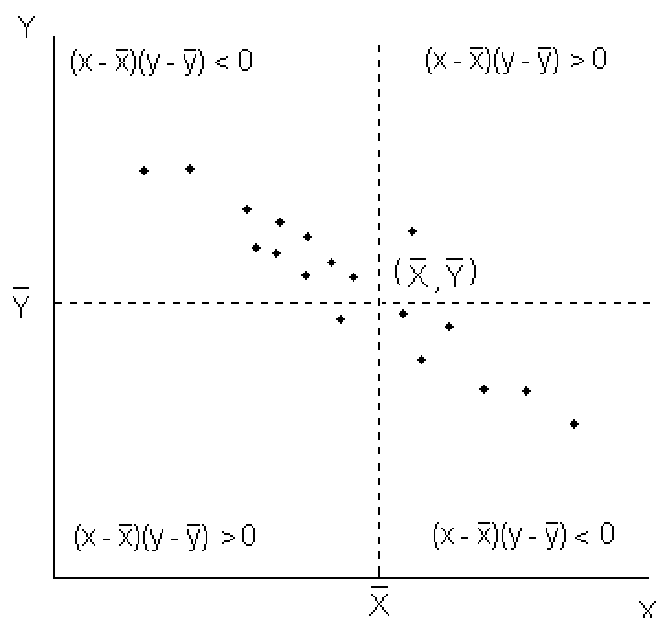
$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}).(Y_i - \overline{Y})}{n-1}$$

Fórmula abreviada: $S_{XY}=\frac{\sum_{i=1}^{n} X_i . Y_i - n\bar{X}\,\bar{Y}}{n-1}$

Propiedades:

- Puede tomar cualquier valor real ($-\infty \leq S_{XY} \leq +\infty$).
- Indica la presencia de asociación lineal y su signo:

$$S_{XY}=\begin{cases} <0 & \text{Asociación lineal negativa} \\ =0 & \text{No existe asociación lineal} \\ >0 & \text{Asociación lineal positiva} \end{cases}$$



- Si dos variables X e Y son estadísticamente independientes, la covarianza es 0. Pero si S_{XY} = 0 **no** implica que las variables sean independientes.
- La covarianza queda afectada por los cambios de escala, pero no por los cambios de origen:
  $X'=a+bX \quad Y'=c+dY \quad \Rightarrow \quad S_{X'Y'}=bdS_{XY}$
  En consecuencia le afectan los cambios de unidades de medida.

Inconvenientes:

- No está acotada.
- No es adimensional. Tiene unidades de medida.

COEFICIENTE DE CORRELACIÓN LINEAL, r_{XY}

Su definición solventa los inconvenientes de la covarianza.

$$r_{XY}=\frac{S_{XY}}{S_X\,S_Y}$$

Indica el grado de asociación lineal que existe entre variables cuantitativas.

Propiedades:

- Está acotado: siempre toma valores comprendidos entre -1 y +1

  $(-1\leq r_{XY}\leq +1)$.

$$r_{XY}=\begin{cases}=-1 & \text{Asociación lineal negativa perfecta}\\ =0 & \text{No existe asociación lineal}\\ =+1 & \text{Asociación lineal positiva perfecta}\end{cases}$$

- Indica el grado de asociación lineal y su signo.

$$r_{XY}=\begin{cases}r_{XY}\approx\pm1 & \text{Alto grado de asociación lineal}\\ r_{XY}\approx 0 & \text{Débil grado de asociación lineal}\end{cases}$$

- Las transformaciones lineales sólo le afectan si hay cambio de signo en el cambio de escala.
- El coeficiente de correlación lineal coincide con la covarianza de las variables estandarizadas.

RESUMEN DEL ANÁLISIS DESCRIPTIVO BIDIMENSIONAL

Vector de Medias: $\left[\bar{X},\,\bar{Y}\right]$ Centro de gravedad de la nube de puntos.

Matriz de Varianzas y Covarianzas: $S^2=\begin{bmatrix}S_X^2 & S_{XY}\\ S_{XY} & S_Y^2\end{bmatrix}$ Dispersión de la nube de puntos.

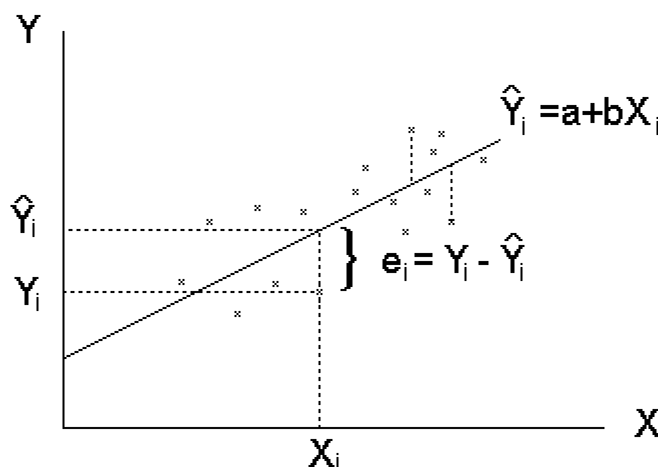
Matriz de Coeficientes de Correlación: $r=\begin{bmatrix} 1 & r_{XY} \\ r_{XY} & 1 \end{bmatrix}$ Grado de asociación lineal.

REGRESIÓN LINEAL

La RECTA DE REGRESIÓN LINEAL (MRLS) modeliza la relación de causalidad de una variable Y con respecto a otra X, de forma que el comportamiento de una viene explicado de forma LINEAL por la otra.

X es la variable independiente o exógena que explica el comportamiento de Y, variable dependiente o endógena.

La recta de regresión lineal $\hat{Y}_i = a+bX_i$ será aquella que mejor se ajuste a la nube de puntos.



Siendo: a: ordenada en el origen
b: pendiente de la recta
e_i: error de predicción o residuo

A estos efectos, utilizando el método de Mínimos Cuadrados Ordinarios (MCO) se obtienen los valores **a** y **b**, tales que determinan la recta que minimiza la variación o dispersión de las observaciones a su alrededor, $\min\sum_{i=1}^{n} e_i^2 = \min\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$$b = \frac{S_{XY}}{S_X^2}$$

$$a = \bar{Y} - b\bar{X}$$

- Los valores $\hat{Y}_i$ recogidos en la recta son estimaciones de los promedios de Y para valores concretos de X.
- El valor **b**, pendiente de la recta, recoge una estimación de la variación de la variable Y por cada incremento unitario de X.
- El valor **a**, ordenada en el origen, recoge el valor ajustado de Y (estimación) suponiendo nulo el valor de X.

Características:

- La recta de regresión siempre pasa por el punto ($\overline{X}, \overline{Y}$).
- La pendiente (**b**) presenta el mismo signo que la covarianza entre X e Y (S_{XY}).
- El coeficiente de correlación lineal, rxy, está directamente relacionado con **b**: $r_{XY} = (S_X/S_Y)b$ y siempre presentan el mismo signo.
- la recta de regresión de Y sobre X (Y= a+bX), en general, no presenta la misma solución que la regresión de X sobre Y (X= c+dY).

Coeficiente de Determinación

Mide el grado de Bondad del Ajuste indicando si el modelo (la recta de regresión) es bueno para explicar la relación de causalidad entre las dos variables, es decir, si la variable independiente X explica el comportamiento de la variable dependiente Y.

Se define como el cociente entre la variación de Y explicada por X (la recogida por la recta de regresión) y la variación total observada en Y: R^2 = VE/VT. Por lo tanto, cuantifica el tanto por uno de la variación observada en Y que queda explicada por la recta ajustada

Toma valores acotados entre 0 y 1: $0 \leq R^2 \leq 1$.

- $R^2 = 1$ significa que el ajuste es perfecto (la nube de puntos está sobre la recta),
- $R^2 = 0$ entonces es que no existe relación lineal entre las dos variables. Es decir, X no explica de forma lineal el comportamiento de Y, por lo tanto el modelo especificado no es el adecuado.

Se demuestra que $R^2 = r_{XY}^2$.

ACTIVIDADES

Actividad 6_1

La base de datos **Ejercicio61.rda** contiene información correspondiente a 204 modelos de automóviles de las variables:

CO_2 Emisiones (en gCO_2/km)
Cilindrada (en cm3)
Consumo (en l/100km)

Utilizando el programa R-Commander halle la matriz de correlación e indique:

a) ¿Cuál es la variable más correlacionadas con CO2?
b) ¿Cree que alguna de estas correlaciones responde a una relación causal?
c) Ajuste la recta de regresión de CO2 sobre la variable que le parezca más adecuada para explicar su comportamiento.
d) Comente qué porcentaje de la variación total observada en CO2 queda retenida por la recta ajustada en el apartado anterior.
e) Obtenga el diagrama de dispersión para CO2 sobre la variable explicativa.
f) Realice una predicción de la emisión de CO2 de un vehículo con un consumo de 6 l/100Km y 1890 cm^3 de cilindrada.

Instrucciones R-Commander:

Matriz de correlación: Estadísticos ▶ Resumenes ▶ Matriz de correlaciones
Regresión: Estadísticos ▶ Ajuste de modelos ▶ Regresión lineal
Gráficos: Gráficas ▶ Diagrama de dispersión

a) Variables más correlaciona con CO2

Matriz de coeficientes de correlación

```
           Cilindrada       CO2   Consumo
Cilindrada  1.0000000 0.7361019 0.6316759
CO2         0.7361019 1.0000000 0.9692480
Consumo     0.6316759 0.9692480 1.0000000
```

La variable más correlacionada con CO2 es Consumo con r_{XY} = 0,969248

b) ¿Cree que alguna de estas correlaciones responde a una relación causal?

Dada la naturaleza de las variables analizadas la relación entre Emisión de CO2 y Consumo o Cilindrada son relaciones de tipo causal: cuanto mayor es el Consumo o la Cilindrada del coche mayor será la emisión de CO2.

c) Recta de regresión ajustada por MCO.

De acuerdo con el apartado a) es preferible explicar la Emisión de CO2 con la variable Consumo. El resultado es:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.4670     2.4121   2.266   0.0245 *
Consumo      23.4319     0.4186  55.979   <2e-16 ***

Residual standard error: 8.328 on 202 degrees of freedom
Multiple R-squared: 0.9394, Adjusted R-squared: 0.9391
F-statistic:  3134 on 1 and 202 DF,  p-value: < 2.2e-16
```

La recta ajustada es:

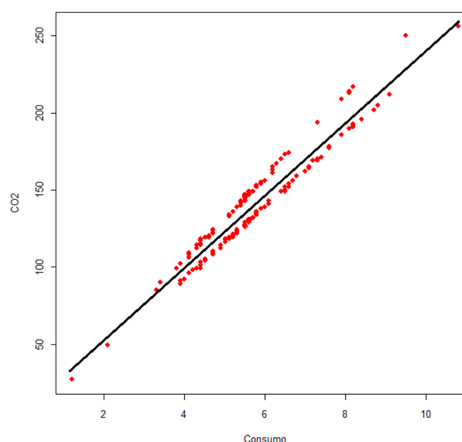
CO2= 5,467 + 23,4319 Consumo

Se estima que en promedio la emisión de CO2 incrementa 23,4319 gr/km cuando el consumo del vehículo aumenta 1 litro/100km.

d) Comente qué porcentaje de la variación total observada en CO2 queda retenida por la recta ajustada en el apartado anterior.

El coeficiente de determinación es R^2 = 0,9394. El 93,94% de la variación total de CO2 en la muestra queda retenido por la recta ajustada, en consecuencia el 6,06% no queda explicado por la relación con Consumo.

e) Obtenga el diagrama de dispersión para X6 sobre X5.



f) Predicción de la emisión de CO_2 de un vehículo con un consumo de 6 l/100Km y 1890 cm^3 de cilindrada.

$$CO_2 = 5{,}467 + 23{,}4319\ \text{Consumo} = 5{,}467 + 23{,}4319\ (6) = 146{,}0584\ \text{gr}$$

Actividad 6_2

A partir de una muestra de 40 franquicias, se desea especificar un modelo de regresión lineal simple que explique el número de unidades vendidas mensualmente de un determinado artículo para el hogar (NVENT).

La información recogida es la siguiente:

	NVENT	VEND	GPU	PRE
1	262	14	600	200
2	261	15	1000	210
3	206	14	1481	211
4	204	17	1237	212
5	196	15	1248	230
6	169	10	900	232
7	178	16	712	244
8	201	16	1000	245
9	233	13	1063	245
10	164	12	555	250
11	187	10	881	250
12	180	12	1045	259
13	165	12	805	261
14	128	8	800	265
15	188	12	1259	265
16	186	12	1237	266
17	130	10	1050	287
18	167	10	1053	289
19	171	10	1214	295
20	138	8	1365	298

	NVENT	VEND	GPU	PRE
21	139	8	1050	299
22	182	15	1401	299
23	89	8	820	300
24	134	10	1262	309
25	91	15	842	310
26	102	7	1270	310
27	159	13	1336	313
28	118	11	1250	317
29	140	9	1120	324
30	139	9	1039	336
31	107	10	974	350
32	118	15	918	354
33	163	10	1337	356
34	81	9	721	370
35	76	6	634	381
36	86	6	600	384
37	120	10	1280	386
38	75	10	1200	390
39	117	10	1310	390
40	52	10	1000	400

VEND: Número de vendedores.
GPU: Gasto mensual en publicidad (en Euros)
PRE: Precio del artículo más gastos de envío (en Euros)

Los resultados del análisis con R Commander han sido los siguientes:

```
> cor(Ej3_3[,c("GPU","NVENT","PRE","VEND")], use="complete.obs")
             GPU      NVENT         PRE       VEND
GPU   1.00000000  0.1397043  0.03732601  0.1321667
NVENT 0.13970431  1.0000000 -0.84447473  0.6277927
PRE   0.03732601 -0.8444747  1.00000000 -0.5692014
VEND  0.13216670  0.6277927 -0.56920136  1.0000000

> numSummary(Ej3_3[,c("GPU", "NVENT", "PRE", "VEND")], statistics=c("mean",
+   "sd"), quantiles=c(0,.25,.5,.75,1))
          mean         sd  n
GPU   1046.725 246.993511 40
NVENT  150.050  50.046619 40
PRE    297.300  57.197005 40
VEND    11.175   2.872393 40

> RegModel.1 <- lm(NVENT~GPU, data=vent)
> summary(RegModel.1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 120.42000   34.98080   3.442  0.00142 **
GPU           0.02831    0.03255   0.870  0.38991

Multiple R-squared:  0.01952


> summary(RegModel.2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 369.72621   23.00592  16.071  < 2e-16 ***
PRE          -0.73890    0.07602  -9.719 7.48e-12 ***

Multiple R-squared:  0.7131

> RegModel.3 <- lm(NVENT~VEND, data=vent)
> summary(RegModel.3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    27.82      25.36   1.097     0.28
VEND           10.94       2.20   4.972 1.45e-05 ***

Multiple R-squared:  0.3941
```

Se pide:

a) ¿Qué tanto por ciento de la variación total de NVENT(Y) queda explicado por cada uno de los siguientes modelos de RLS?

$Y = a + b\ (VEND)$
$Y = a + b\ (GPU)$
$Y = a + b\ (PRE)$

b) Indique cuál de las tres variables proporciona un modelo RLS con mayor capacidad explicativa del comportamiento de la variable NVENT.
c) Razone cuál será el signo de la pendiente de la recta seleccionada.
d) Halle la recta de regresión más adecuada.
e) ¿Cómo repercute sobre el número de unidades vendidas un incremento de 1 Euro en la variable PRE?
f) ¿Cuántas unidades espera vender una franquicia para la cual: VEND=11, GPU=1050 y PRE= 300 Euros?
g) Obtenga la predicción de Y para otra franquicia con: VEND=11, GPU=1050 y PRE= 450 Euros. Compare la fiabilidad de esta predicción con la del apartado anterior.

a) Análisis de la correlación de NVENT con VEND, GPU y PRE

	GPU	NVENT	PRE	VEND
GPU	1.00000000	0.1397043	0.03732601	0.1321667
NVENT	0.13970431	1.0000000	-0.84447473	0.6277927
PRE	0.03732601	-0.8444747	1.00000000	-0.5692014
VEND	0.13216670	0.6277927	-0.56920136	1.0000000

El coeficiente de determinación del la recta Y = a + b (VEND) es
$R^2=r^2_{Y,VEND}=0,628^2 = 0,3943$

El coeficiente de determinación del la recta Y = a + b (GPU) es
$R^2=r^2_{Y,GPU}=0,140^2 = 0,0196$

El coeficiente de determinación del la recta Y = a + b (PRE) es
$R^2=r^2_{Y,PRE}= (-0,844)^2 = 0,7123$

Estas rectas explican el 39,43%, 1,96% y 71,23% de la variación total de Y, respectivamente.

b) La variable con mayor capacidad explicativa del comportamiento de Y (NVENT)

La variable más correlacionada con NVENT es PRE.

c) Signo de la recta de regresión

La recta seleccionada es Y = a + b (PRE). Como el coeficiente de correlación entre estas variable es negativo, la pendiente de la recta también será negativa.

d) Recta de regresión muestral

La recta ajustada es NVENT = 369,726– 0,7389 PRE

e) Si PRE se incrementa en 1 Euro,

el número de unidades vendidas mensualmente se espera que disminuya en 0,7389 unidades.

f) Predicción para una franquicia con PRE = 300 Euros la predicción es

NVENT = 369,726– 0,7389 (300) = 148,0 unidades

g) Para una franquicia con PRE = 450 la predicción es:

NVENT = 369,726– 0,7389 (450) = 37,2 unidades
La fiabilidad de esta predicción es menor que la del apartado anterior porque se genera para un valor de la variable explicativa PRE que está fuera del recorrido de esta variable en la muestra.

Actividad 6_3

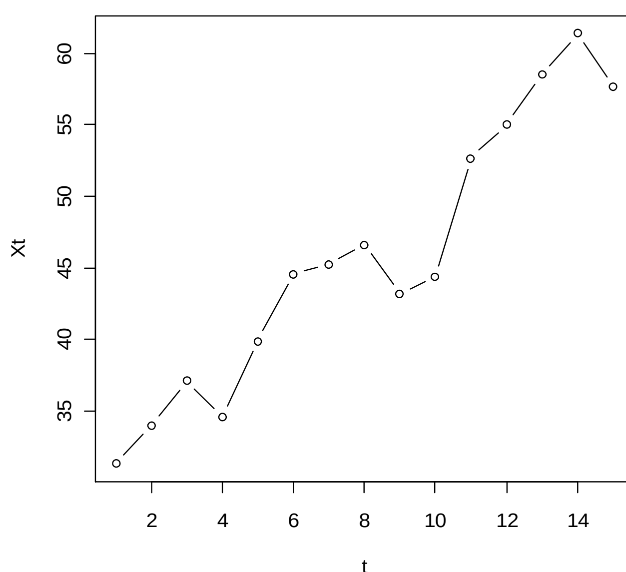
La siguiente serie recoge la evolución anual de la variable X = 'Importe global de la participación en los beneficios', en miles de Euros, de los vendedores de un concesionario de automóviles.

Año	1995	1996	1997	1998	1999	2000	2001	2002
t	1	2	3	4	5	6	7	8
Xt	31,29	34,01	37,12	34,6	39,88	44,56	45,24	46,62

Año	2003	2004	2005	2006	2007	2008	2009
t	9	10	11	12	13	14	15
Xt	43,2	44,38	52,66	55,04	58,52	61,4	57,68

a) Ajuste y represente gráficamente la serie.
b) Obtenga las predicciones de la tendencia para los años 2010 y 2011.

a) Representación gráfica



Estimación de la tendencia

A partir de la serie de observaciones se obtienen los siguientes resultados:

$$\sum_{t=1}^{15} X_t = 686{,}2 \quad \sum_{t=1}^{15} t = 120 \quad \sum_{t=1}^{15} t^2 = 1240 \quad \sum_{t=1}^{15} X_t^2 = 32666{,}723 \quad \sum_{t=1}^{15} tX_t = 6063{,}37$$

$$S_{tX} = \frac{\sum_{t=1}^{15} tX_t - 15\,\bar{t}\,\bar{X}}{15-1} = \frac{6063,37 - 15\frac{120}{15}\frac{686,2}{15}}{14} = 40,98$$

$$S_t^2 = \frac{\sum_{t=1}^{15} t^2 - 15\,\bar{t}^2}{15-1} = \frac{1240 - 15\left(\frac{120}{15}\right)^2}{14} = 20,00$$

$$S_X^2 = \frac{\sum_{t=1}^{15} X_t^2 - 15\,\bar{X}^2}{15-1} = \frac{32666,723 - 15\left(\frac{686,2}{15}\right)^2}{14} = 91,09$$

$$b = \frac{S_{tX}}{S_t^2} = \frac{40,98}{20,00} = 2,05 \qquad a = \bar{X} - b\bar{t} = \frac{686,2}{15} - 2,05\frac{120}{15} = 29,35$$

$$\hat{X}_t = 29,35 + 2,05\ t$$

Origen: $t = 1$ en el año 1995

$$R^2 = r_{tX}^2 = \frac{S_{tX}^2}{S_t^2 S_X^2} = \frac{40,98^2}{(20)(91,09)} = 0,922$$

La tendencia recoge el 92,2% de la variación total de X

b) Predicción para los años 2010 y 2011:

Para el año 2010 $t = 16$ $\hat{x}_{2010} = 29,35 + 2,05(16) = 62.15$

Para el año 2011 $t = 17$ $\hat{x}_{2011} = 29,35 + 2,05(17) = 64,20$

Resumen de los resultados con R-Commander

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.3532     1.5040   19.52 5.17e-11 ***
t             2.0492     0.1654   12.39 1.43e-08 ***
---
Multiple R-squared: 0.9219, Adjusted R-squared: 0.9159
```

EJERCICIOS TEMA 6

Ejercicio1. A partir de los siguientes datos correspondientes a 8 empresas de USA sobre las ventas y los beneficios obtenidos en billones de dólares:

Empresa	ventas	beneficios
General Motors	78	1,25
Exxon	69	1,10
Ford	62	0,77
IBM	51	0,56
Mobil	44	0,45
General Electric	35	0,44
AT&T	34	0,44
Texaco	31	0,42

Se pide:

a) Obtenga las medidas de asociación.
b) Si las ventas de todas estas empresas se incrementan en un 5% y los beneficios en 0,5 billones de dólares, ¿cuál será la covarianza de las variables transformadas? ¿y el coeficiente de correlación lineal?

Ejercicio 2. Un concesionario, para analizar la aceptación de dos nuevos modelos de motocicleta ha observado durante los 25 días laborables del último mes las unidades vendidas:

X= Unidades vendidas del modelo A

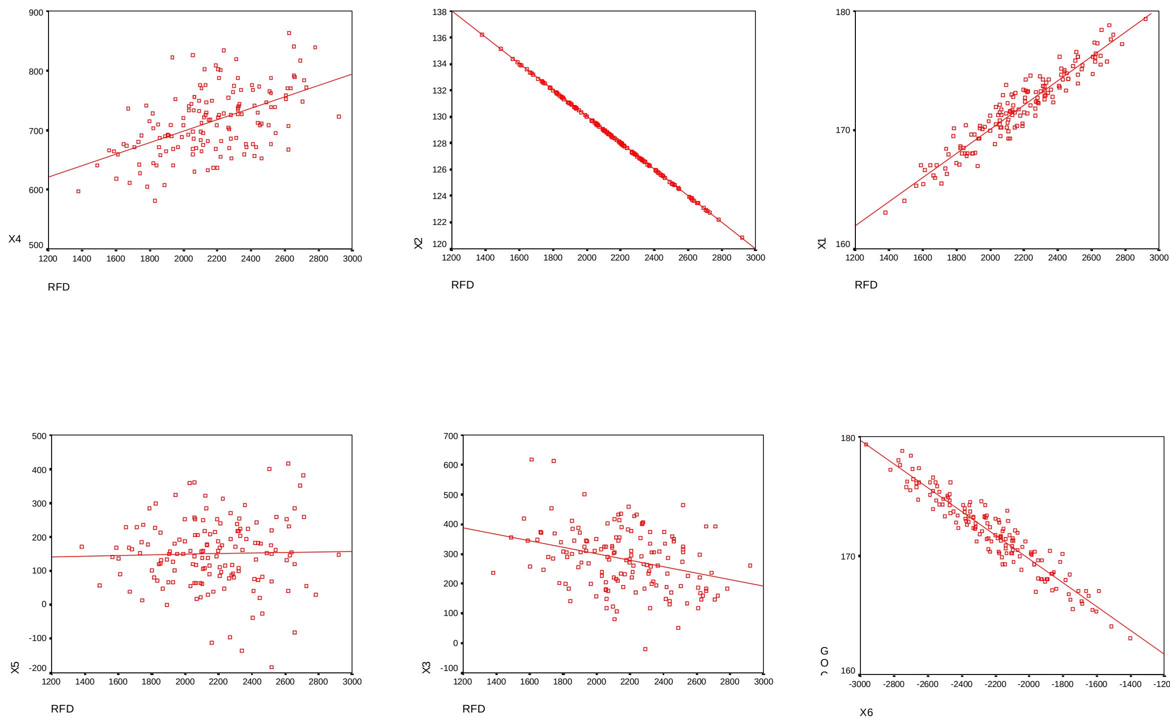
Y= Unidades vendidas del modelo B

X Modelo A	Y Modelo B	Núm. días
0	3	1
1	1	5
2	1	10
3	2	9

Se pide:

a) Vector de medias y matriz de varianzas y covarianza.
b) Coeficiente de correlación lineal y su interpretación.
c) Si el número de motos vendidas de cada modelo en cada uno de los días observados fuera el doble cuál sería la covarianza y el coeficiente de correlación entre las unidades vendidas de los dos modelos.

Ejercicio3. Los siguientes diagramas de dispersión corresponden a observaciones conjuntas de 6 pares de variables (X,Y):



Asigne a cada uno de los diagramas el valor del coeficiente de correlación que le parezca más adecuado: r=0,98; r=0,03; r=-0,42; r=-0,95; r=-1; r=0,32.

Ejercicio 4. A partir de una muestra de 100 observaciones referente a las variables:

Y = Saldo de las Imposiciones en Cajas de Ahorros
X = Renta Familiar Disponible

se han obtenido los siguientes resultados:

$$\overline{X} = 4{,}65 \quad \overline{Y} = 1{,}55 \quad S_X^2 = 5{,}48 \quad S_Y^2 = 1{,}04 \quad S_{XY} = 2{,}13$$

Se pide:

a) Determine el grado de asociación lineal entre estas variables.
b) Obtenga la ecuación de regresión lineal que explica el Saldo de las Imposiciones en función de la Renta Familiar.
c) Indique el porcentaje de variación observado en Y explicado por el ajuste anterior

Ejercicio 5. La siguiente tabla recoge la edad (X) y la presión sanguínea máxima (Y) de un grupo de 10 mujeres:

Edad	56	42	72	36	63	47	55	49	38	42
Presión	14,8	12,6	15,9	11,8	14,9	13	15,1	14,2	11,4	14,1

a) Calcule el coeficiente de correlación lineal entre las variables anteriores y comente el resultado obtenido.
b) Determine la recta de regresión de Y sobre X justificando la adecuación de un ajuste lineal. Interprete los coeficientes.
c) Valore la bondad del ajuste.
d) Obtenga las siguientes predicciones, únicamente en los casos que tenga sentido hacerlo:

Presión sanguínea de una mujer de 51 años.
Presión sanguínea de una niña de 10 años.
Presión sanguínea de un hombre de 54 años.

Ejercicio 6. A fin de analizar si la lluvia caída puede ser explicativa de la calidad del vino, se han observado en 60 muestras la calificación del vino en grados (Y) y la lluvia anual registrada (X), obteniendo los siguientes resultados:

$$\sum_{i=1}^{60} X_i=32400 \quad \sum_{i=1}^{60} Y_i=640 \quad \sum_{i=1}^{60} X_i^2=21280000$$
$$\sum_{i=1}^{60} Y_i^2=7088 \quad \sum_{i=1}^{60} X_iY_i=321600$$

a) Obtenga la regresión lineal que recoge el comportamiento de la calidad del vino en función de la lluvia.
b) Indique el porcentaje de variación observada en la graduación del vino que viene explicada por la recta de regresión obtenida.
c) ¿En cuánto se puede estimar la variación de la graduación del vino si la lluvia caída aumenta 10 unidades?
d) ¿Qué predicción haría de la graduación del vino obtenido de una cosecha en un año en que la lluvia registrada fuera de 950?

Ejercicio 7. El vector de medias y la matriz de varianzas y covarianzas de las variables X_1 = Ingreso semanal y X_2 = Gasto semanal (en u.m.) observadas sobre un conjunto de 100 familias son:

$$\bar{X} = \begin{bmatrix} 1250 & 1050 \end{bmatrix}' \quad S^2 = \begin{bmatrix} 1200 & 900 \\ & 4000 \end{bmatrix}$$

a) Calcule el coeficiente de correlación lineal.
b) Obtenga el ajuste lineal del Gasto semanal en función del Ingreso.
c) ¿Qué porcentaje de variación del Gasto no queda explicado por el ajuste anterior?
d) ¿Qué predicción del Gasto semanal haría para una familia con 1100 u.m. de Ingreso semanal? Comente la validez del resultado anterior.

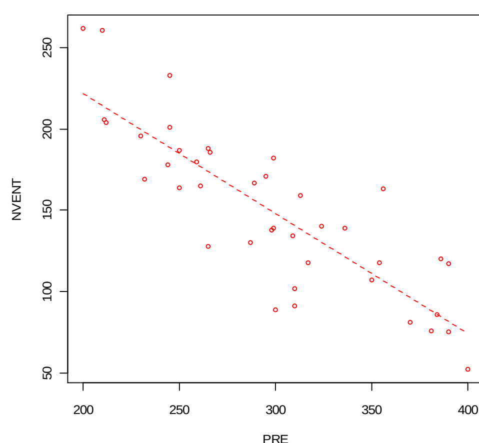
Ejercicio 8. Suponga que ha observado simultáneamente el precio X y la demanda Y de un determinado producto, obteniendo:

X	1	2	3	4	5	6	7	10	10	12
Y	12	10	10	8	6	5	7	6	4	2

$$\sum_{i=1}^{10}(X_i-\bar{X})^2=124 \quad \sum_{i=1}^{10}(Y_i-\bar{Y})^2=84 \quad \sum_{i=1}^{10}(X_i-\bar{X})(Y_i-\bar{Y})=-93$$

a) Obtenga el ajuste lineal de la demanda en función del precio.
b) Determine la bondad del ajuste.
c) Compruebe los resultados con los obtenidos con el R-Commander.
d) Si el precio se incrementa en 1 unidad, ¿cuál es el incremento esperado en la demanda?
e) Obtenga la predicción de la demanda esperada si el precio es 9.

Ejercicio 9. El siguiente diagrama de dispersión corresponde a la distribución de frecuencias conjuntas de las variables NVENT = "Nº de unidades vendidas mensualmente de determinado artículo" y PRE = "precio del artículo en Euros·".



Razone que signo y magnitud aproximada presentarán la pendiente y el coeficiente de determinación de la recta de regresión ajustada.

Ejercicio 10. Sobre una muestra de 61 pisos vendidos en el área metropolitana de Barcelona durante el último trimestre del año 2011 se han observado las variables:

Preu.de.venda = precio pagado por el comprador (Euros)
Preu.inicial = primer precio ofrecido al comprador (Euros)
Metres.quadrats = superficie en m^2

Con el programa R-Commander se han obtenido los siguientes resultados:

Análisis descriptivo unidimensional:

```
                     mean          sd  n
Metres.quadrats      65.06721    12.13026 61
Preu.de.venda    170109.67803 35628.50863 61
Preu.inicial     181766.75115 36727.83416 61
```

Matriz de correlación:

```
                Metres.quadrats Preu.de.venda Preu.inicial
Metres.quadrats       1.0000000     0.9510897    0.9668259
Preu.de.venda         0.9510897     1.0000000    0.9828272
Preu.inicial          0.9668259     0.9828272    1.0000000
```

Para explicar el comportamiento de la variable Preu.de.venda se proponen las siguientes rectas de regresión lineal:

Recta I

```
Coefficients:
                Estimate
(Intercept)     -11655.7
Metres.quadrats   2793.5
```

Recta II

```
Coefficients:
               Estimate
(Intercept)  -3188.4840
Preu.inicial     0.9534
```

Se pide:

a) Razone porqué la recta II tiene mayor capacidad explicativa del comportamiento de Preu.de.venda que la recta I e indique qué porcentaje de la variación total de Preu.de.venda queda explicado por esta recta.
b) Con respecto a la recta II, indique en cuánto repercute un aumento de 1000 Euros en el Preu.inicial sobre el Preu.de.venda esperado.
c) Estime cuál será el Preu.de.venda de un piso cuyo precio inicial es de 150000 Euros.
d) Estime cuál será el Preu.de.venda de un piso cuyo precio inicial es de 270000 Euros.
e) ¿Cuál de las dos predicciones es más fiable? Razone la respuesta.

Tema 7. INTRODUCCIÓN A LA TEORÍA DE LA PROBABILIDAD

Experimento aleatorio. Probabilidad: axiomática y propiedades
Probabilidad condicionada
Teorema de la intersección. Independencia de sucesos
Teorema de la probabilidad total. Teorema de Bayes

La probabilidad se ocupa de medir o determinar cuantitativamente la posibilidad de que un suceso o experimento produzca un determinado resultado.

Toda medida de probabilidad, como posibilidad de ocurrencia de un suceso, debe cumplir la Axiomática de Kolmogorov y las propiedades que de ella se deducen.

La asignación de la probabilidad a los distintos sucesos se determina aplicando alguno de los siguientes criterios: Teoría clásica o Regla de Laplace, que exige equiprobabilidad de los resultados elementales, la teoría frecuencialista que exige repetición del experimento, y la teoría subjetivista, cuando no se dan los requisitos anteriores.

La información adicional de la ocurrencia de un suceso nos lleva al concepto de PROBABILIDAD CONDICIONADA y al TEOREMA DE LA INTERSECCIÓN.

Si la probabilidad de un suceso no se ve modificada por el hecho de haberse verificado otro, entonces se concluye que ambos sucesos son INDEPENDIENTES.

Dado un conjunto de sucesos que inducen una partición en el espacio referencial de un experimento y cuyas probabilidades son conocidas, el TEOREMA DE LA PROBABILIDAD TOTAL permite calcular la probabilidad de cualquier otro suceso que se presenta siempre acompañado por uno de aquellos.

El TEOREMA DE BAYES, base de la concepción bayesiana de la Estadística, es de enorme relevancia puesto que vincula la probabilidad de A dado B con la probabilidad de B dado A. Permite reasignar probabilidades establecidas a priori a partir de una información adicional.

EXPERIMENTO ALEATORIO.

La estadística es la ciencia empírica que estudia los fenómenos que dependen de azar. Estos fenómenos aleatorios están asociados a experimentos que se pueden repetir de forma ilimitada y presentan resultados imprevisibles aunque se realicen en las mismas condiciones. Los experimentos aleatorios, a pesar de estos resultados imprevisibles, se caracterizan porque presentan una pauta de comportamiento o regularidad estadística a largo plazo que, como se verá, puede, generalmente, modelizar con alguno de los modelos de probabilidad que se estudiarán más adelante.

Espacio muestral

El espacio muestral o espacio referencial, E, es el conjunto de todos los resultados posibles de un experimento aleatorio.

Suceso aleatorio

Dado un espacio referencial, se define como suceso aleatorio cualquier subconjunto de dicho espacio referencial. El suceso que contiene un solo resultado de E se llama suceso elemental o elemento muestral.

Ejemplo. Una caja contiene 1 bola blanca (B), 1 bola roja (R) y 1 bola negra (N). El experimento consiste en extraer dos bolas con devolución y observar la secuencia de colores obtenida. El espacio muestral o referencial es:

E = {BB, BR, BN, RB, RR, RN, NB, NR, NN}

Sobre este espacio pueden definirse, entre otros, los siguientes sucesos:

A_1 = 'Dos bolas del mismo color' = {BB, RR, NN}
A_2 = 'Dos bolas de distinto color' = {BR, BN, RB, RN, NB, NR}
A_3 = 'Por lo menos una bola blanca' = {BB, BR, BN, RB, NB}
A_4 = 'Exactamente una bola blanca' = {BR, BN, RB, NB}
A_5 = 'Ninguna bola blanca' = {RR, RN, NR, NN}
A_6 = 'Dos bolas blancas' = {BB}

Relaciones entre sucesos

I) **Suceso complementario**- Dado un suceso A, el suceso complementario de A, $\overline{A}$, es aquel que contiene todos los resultados del experimento que no están contenidos en A; es decir, $\overline{A}$ es el suceso que ocurre cuando no ocurre A.

Por ejemplo:

$\overline{A}_1$ = {BR, BN, RB, RN, NB, NR}

$\bar{A}_2 = A_1$
$\bar{A}_3 = \{RR, RN, NR, NN\}$
$\bar{A}_4 = \{BB, RR, RN, NR, NN\}$
$\bar{A}_5 = A_3$
$\bar{A}_6 = \{BR, BN, RB, RR, RN, NB, NR, NN\}$

II) **Suceso unión**- Dados dos sucesos A_i y A_j el suceso unión $(A_i \cup A_j)$ es aquel que contiene todos los resultados que pertenecen a A_i, a A_j o a ambos.

Por ejemplo:
$(A_1 \cup A_3) = \{BB, BR, BN, RB, RR, NB, NN\}$
$(A_1 \cup A_5) = \{BB, RR, RN, NR, NN\}$
$(A_4 \cup A_6) = \{BB, BR, BN, RB, NB\} = A_3$

- Si $(A_i \cup A_j) = A_i$ se dice que A_j está contenido en A_i; es decir todos los resultados que pertenecen a A_j pertenecen también a A_i de forma que si ocurre A_j ocurre también A_i.

Por ejemplo: $(A_1 \cup A_6) = \{BB, RR, NN\} = A_1$ A_6 está contenido en A_1 $(A_6 \subset A_1)$

- La unión de un suceso cualquiera A y su complementario $\bar{A}$ es E.

Por ejemplo: $(A_1 \cup A_2) = \{BB, RR, NN, BR, BN, RB, RN, NB, NR\} = E$

- La unión puede generalizarse para más de dos sucesos.

Por ejemplo: $(A_1 \cup A_5 \cup A_6) = \{BB, RR, RN, NR, NN\}$

III) **Suceso intersección**- Dados dos sucesos A_i y A_j el suceso intersección $(A_i \cap A_j)$ es aquel que contiene todos los resultados que pertenecen a A_i y a A_j simultáneamente.

Por ejemplo:
$(A_1 \cap A_3) = \{BB\}$
$(A_1 \cap A_5) = \{RR, NN\}$
$(A_1 \cap A_6) = \{BB\}$

- Si un suceso A_j está contenido en A_i, $(A_i \cap A_j) = A_j$.

Por ejemplo: A_6 está contenido en A_1 $(A_6 \subset A_1) \Rightarrow (A_1 \cap A_6) = \{BB\} = A_6$

- Cuando dos sucesos A_i y A_j son tales que $(A_i \cap A_j) = \phi$ se dice que A_i y A_j son sucesos incompatibles o **mutuamente excluyentes**.

Por ejemplo:

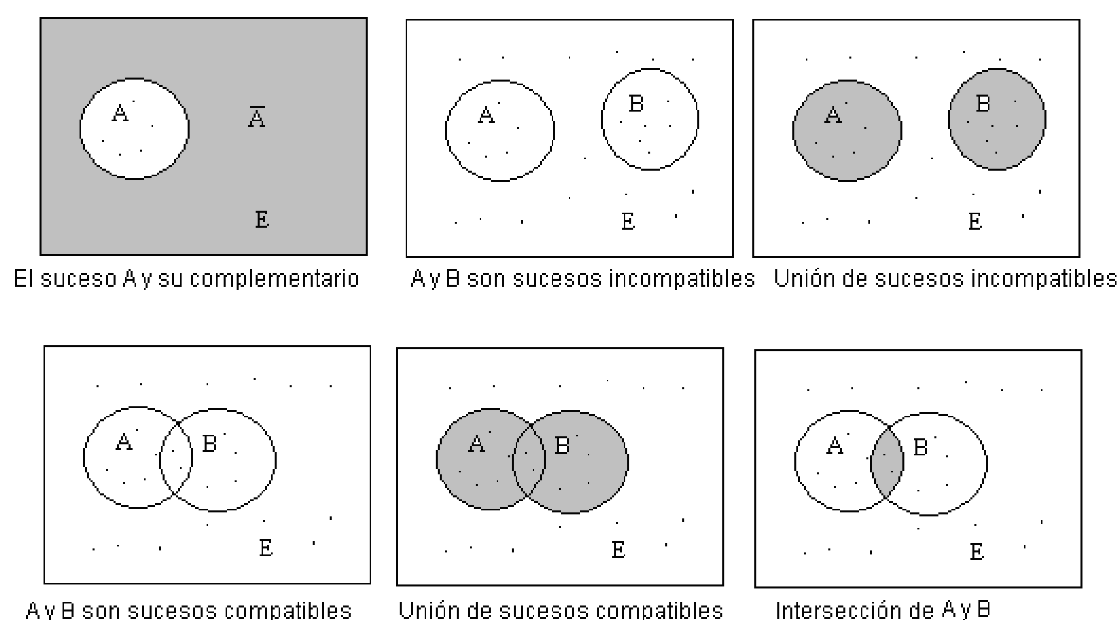
$(A_1 \cap A_4) = \phi$ A_1 y A_4 son incompatibles

$(A_4 \cap A_6) = \phi$ A_4 y A_6 son incompatibles

- La intersección de un suceso A y su complementario $\overline{A}$ es igual al conjunto vacío, por tanto A y $\overline{A}$ son sucesos incompatibles.
- La intersección puede generalizarse para más de dos sucesos.

Por ejemplo: $(A_1 \cap A_5 \cap A_6) = \{BB\}$

Diagramas de Venn



PROBABILIDAD. AXIOMÁTICA Y PROPIEDADES

Dado un espacio referencial E se dice que P es una función de probabilidad definida en E si a cualquier suceso, A, de E le hace corresponder un número real, P(A), que mide el grado de posibilidad de la ocurrencia de A, y verifica los siguientes axiomas:

Axioma I $P(A) \geq 0$

Axioma II $P(E) = 1$

Axioma III   Si A1, A2, A3, ... es una sucesión numerable de sucesos mutuamente excluyentes, la probabilidad de su unión es:

$P(A_1 \cup A_2 \cup A_3 \cup ...) = P(A1) + P(A2) + P(A3) + ...$

De los axiomas se deducen las siguientes propiedades:

1. $P(\phi) = 0$
2. $0 \leq P(A) \leq 1 \qquad \forall A \in E$
3. $P(\bar{A}) = 1 - P(A)$
4. Si $A \subset B$, entonces $P(A) \leq P(B)$
5. Ley aditiva: si A y B son dos sucesos cualesquiera, la probabilidad de su unión es:

   $$(A \cup B) = P(A) + P(B) - P(A \cap B)$$

   Esta propiedad puede generalizarse a tres o más sucesos, por ejemplo:

   $$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Asignación de probabilidad a un suceso

1. Teoría Clásica: Regla de Laplace

Si un espacio referencial, E, contiene un número finito de resultados y éstos son igualmente probables, la probabilidad de un suceso cualquiera A es:

$$P(A) = \frac{\text{Resultados favorables a A}}{\text{Total de resultados posibles}}$$

2. Teoría Frecuencialista

La probabilidad de un suceso A es el límite de la frecuencia relativa de A (n_A/n) cuando el número de experimentos tiende a infinito.

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

3. Teoría Subjetivista

Esta teoría interpreta la probabilidad como el grado de convencimiento subjetivo que cada individuo puede tener en relación a la ocurrencia de un determinado suceso.

Ejemplo: Determinación de la probabilidad de un suceso. Regla de Laplace.
Se han definido los siguientes sucesos sobre el espacio muestral

E= {BB, BR, BN, RB, RR, RN, NB, NR, NN}

A_1 = {BB, RR, NN} — $P(A_1) = 3/9$
A_2 = {BR, BN, RB, RN, NB, NR} — $P(A_2) = 6/9$
A_3 = {BB, BR, BN, RB, NB} — $P(A_3) = 5/9$
A_4 = {BR, BN, RB, NB} — $P(A_4) = 4/9$
A_5 = {RR, RN, NR, NN} — $P(A_5) = 4/9$
A_6 = {BB} — $P(A_6) = 1/9$

$(A_1 \cap A_3)$ = {BB} — $P(A_1 \cap A_3) = 1/9$
$(A_1 \cap A_5)$ = {RR, NN} — $P(A_1 \cap A_5) = 2/9$
$(A_1 \cap A_4) = \phi$ — $P(A_1 \cap A_4) = 0$
$(A_1 \cap A_3 \cap A_6)$ = {BB} — $P(A_1 \cap A_3 \cap A_6) = 1/9$

Probabilidad de la unión de sucesos
$(A_1 \cup A_3)$ = {BB, BR, BN, RB, RR, NB, NN} — $P(A_1 \cup A_3) = 7/9$
A_1 y A_3 no son incompatibles. Por la propiedad 3:
$P(A_1 \cup A_3) = P(A_1) + P(A_3) - P(A_1 \cap A_3) = \frac{3}{9}+\frac{5}{9}-\frac{1}{9}=\frac{7}{9}$

$(A_4 \cup A_6)$ = {BB, BR, BN, RB, NB} — $P(A_4 \cup A_6) = 5/9$
A_4 y A_6 son incompatibles. Por el axioma 3:
$P(A_4 \cup A_6) = P(A_4) + P(A_6) = \frac{4}{9}+\frac{1}{9}=\frac{5}{9}$

$(A_1 \cup A_5 \cup A_6)$ = {BB, RR, RN, NR, NN} — $P(A_1 \cup A_5 \cup A_6) = 5/9$
Por la propiedad 3:
$P(A_1 \cup A_5 \cup A_6) = P(A_1)+P(A_5)+P(A_6)-P(A_1 \cap A_5)-P(A_1 \cap A_6)-P(A_5 \cap A_6)+P(A_1 \cap A_5 \cap A_6) =$
$\frac{3}{9}+\frac{4}{9}+\frac{1}{9}-\frac{2}{9}-\frac{1}{9}-0+0=\frac{5}{9}$

Probabilidad del suceso complementario
$\overline{A}_1$ = {BR, BN, RB, RN, NB, NR} — $P(\overline{A}_1)=6/9$
$\overline{A}_6$ = {BR, BN, RB, RR, RN, NB, NR, NN} — $P(\overline{A}_6) = 8/9$
Por la propiedad 2:
$P(\overline{A}_1) = 1 - P(A_1) = 1 - \frac{3}{9}=\frac{6}{9}$ — $P(\overline{A}_6) = 1 - P(A_6) = 1 - \frac{1}{9}=\frac{8}{9}$

PROBABILIDAD CONDICIONADA

Dados dos sucesos, A y B, de un mismo espacio muestral, con probabilidades no nulas, $P(A) \neq 0$ y $P(B) \neq 0$, la probabilidad de que ocurra el suceso A sabiendo que ha ocurrido el suceso B recibe el nombre de probabilidad condicionada de A respecto a B y se define como:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Análogamente, la probabilidad condicionada de B respecto a A es:

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

TEOREMA DE LA INTERSECCIÓN

De la definición de probabilidad condicionada se deduce que la probabilidad del suceso ($A \cap B$) es:

$P(A \cap B) = P(A)\ P(B/A)$ o $P(B \cap A) = P(B)\ P(A/B)$

El teorema se puede generalizar para n sucesos:

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)\ P(A_2/A_1)\ P(A_3/A_1 \cap A_2)\ \ldots.\ P(A_n / \bigcap_{i=1}^{n-1} A_i)$$

INDEPENDENCIA DE SUCESOS

Dos sucesos, A y B, son estocásticamente independientes si la ocurrencia de uno de ellos no modifica la probabilidad de ocurrencia del otro; es decir, si

$P(A/B) = P(A)$ o $P(B/A) = P(B)$

De donde se deduce que:

si A y B son independientes $\Rightarrow P(A \cap B) = P(A)\ P(B)$

Propiedades:

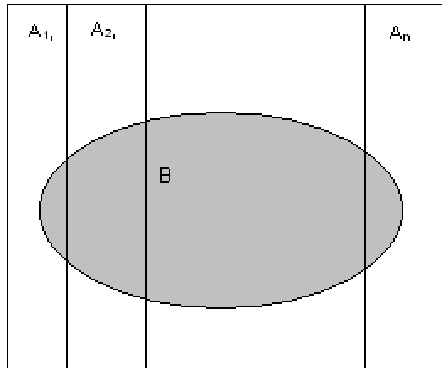
Si dos sucesos A y B son independientes, **no** son incompatibles.
Si dos sucesos A y B son independientes también son independientes $\overline{A}$ y B; A y $\overline{B}$; $\overline{A}$ y $\overline{B}$.

TEOREMA DE LA PROBABILIDAD TOTAL

Si $A_1, A_2, \ldots, A_n$ son sucesos de E con probabilidades no nulas, $P(A_i) \neq 0\ \forall$ i), tales que forman una partición, es decir:

$A_1 \cup A_2 \cup \ldots \cup A_n = E$ (exhaustivos)

$A_i \cap A_j = \phi$ $\forall i \neq j$ (incompatibles)



y B es un suceso de E que ocurre siempre acompañado con uno de los A_i,

la probabilidad de B es:

$P(B) = P[(A_1 \cap B) \cup (A_2 \cap B) \cup \ldots \cup (A_n \cap B)] = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + +\ldots + P(A_n)P(B/A_n)$

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B/A_i)$$

TEOREMA DE BAYES

Sean los sucesos $A_1, A_2, \ldots, A_n$ una partición de E. Asignadas unas probabilidades iniciales $P(A_1), P(A_2), \ldots, P(A_n)$, o probabilidades a priori que reflejan el grado de creencia sobre la ocurrencia de $A_1, A_2, \ldots, A_n$, la realización de un experimento en dicho espacio referencial puede modificar las probabilidades asignadas a priori.

Si el resultado del experimento es el suceso B y se conoce la ocurrencia de B en los sucesos A_i de la partición, es decir, $P(B/A_i)$, esta evidencia experimental permite obtener las nuevas probabilidades de los A_k condicionadas al resultado B, es decir, $P(A_k/B)$ que se denominan probabilidades a posteriori.

$$P(A_k/B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k)P(B/A_k)}{\sum_{i=1}^{n} P(A_i)P(B/A_i)}$$

El teorema de Bayes formula que:

$$P(A_k/B) = P(A_k)\frac{P(B/A_k)}{P(B)}$$

La probabilidad a posteriori es igual a la probabilidad a priori multiplicada por un factor modificativo que depende del resultado del experimento.

EJERCICIOS TEMA 7

Ejercicio 1. Indique cuál es el espacio muestral asociado a los siguientes experimentos aleatorios:

g) De una población con N=7 elementos se extrae una muestra de tamaño n=2 sin devolución.
h) Se observa el número de clientes atendidos por un vendedor hasta que realiza la primera venta.
i) Se observa cada hora la temperatura de una cámara frigorífica que se mantiene a una temperatura de entre 0º y 3º
j) Se elige al azar un aparato eléctrico y se observa el tiempo que transcurre hasta la primera avería.
k) Se echan 5 bolas en 2 cajas, de modo que cada bola tenga la misma probabilidad de caer en cualquiera de las cajas, y se observa el número de bolas que caen en cada caja.

Ejercicio 2. Determine el valor de la probabilidad en cada una de las siguientes situaciones e indique el criterio de asignación utilizado en cada caso:

a) Probabilidad de que un negocio de alimentación tenga éxito si, en términos generales, se estima que por cada negocio de este tipo que fracasa, tres tienen éxito.
b) Probabilidad de obtener una carta de copas al extraer al azar una carta de una baraja española de 48 cartas.
c) Probabilidad de obtener una puntuación total superior a 7 al tirar dos dados.
d) Probabilidad de accidente laboral en un sector industrial sabiendo que en una muestra de 8000 trabajadores, 40 han sufrido un accidente.
e) Probabilidad de que llueva el próximo fin de semana sabiendo que sobre las Azores hay una borrasca.
f) Probabilidad de que llegue con retraso un vuelo del puente aéreo si se sabe que de los 80 vuelos semanales en promedio sólo llegan 2 con retraso.
g) Probabilidad de obtener un número primo al lanzar un dado.
h) Se echan 5 bolas en 2 cajas, de modo que cada bola tenga la misma probabilidad de caer en cualquiera de las cajas, la probabilidad de que en la primera caja caigan exactamente 3 bolas.

Ejercicio 3. A partir de una encuesta sobre hábitos de lectura de la prensa diaria realizada sobre una muestra de 1000 personas se obtiene información sobre los siguientes sucesos:

A = 'lee el diario A'
B = 'lee el diario B'
C = 'lee el diario C'

Resulta que del total de entrevistados:

El 62% lee A
El 92% lee B
El 11% lee C
El 60% lee A y B
El 6% lee A y C
El 11% lee B y C
El 6% lee los tres diarios

1. ¿Son incompatibles los sucesos 'leer el diario A' y 'leer el diario B'?
2. Si se elige al azar un entrevistado, halle:

a) Probabilidad de que no lea A.
b) Probabilidad de que lea A o B.
c) Probabilidad de que lea B o C.
d) Probabilidad de que no lea ni A ni B.
e) Probabilidad de que lea por lo menos uno de los tres diarios.
f) Probabilidad de que no lea ninguno de los tres diarios.
g) Probabilidad de que lea B y no lea C.

Ejercicio 4. Al lanzar dos veces un dado:

a) ¿Cuál es la probabilidad de que la puntuación total sea 8 si se sabe que los dos resultados son diferentes?
b) ¿Cuál es la probabilidad de obtener como máximo 4 puntos si se sabe que ha salido un 2?
c) ¿Cuál es la probabilidad de que haya salido un 2 si se sabe que se ha obtenido más de 4 puntos?

Ejercicio 5. Se sabe que la probabilidad de que se atasque el papel en unas determinadas máquinas fotocopiadoras depende del color del papel. La probabilidad de que la fotocopia se haga en papel blanco y se atasque es 0,05; y la de hacerla en papel de color y atascarse es 0,10. Si el 80% de les fotocopies se hacen con papel blanco:

a) Determine todas las probabilidades conjuntas y marginales de los sucesos 'atascar (si/no)' y 'papel (blanco/color)'.
b) ¿Cuál es la probabilidad de que se atasque una fotocopia que se ha hecho con papel blanco?
c) ¿Cuál es la probabilidad de que el papel sea de color si la fotocopia se ha atascado?
d) Si el papel es de color, ¿cuál es la probabilidad de que la fotocopia no se atasque?

Ejercicio 6. Sobre determinado colectivo se sabe que:

- La edad del 25% es menor de 30 años.
- El 60% de los que tienen menos de 30 años practica algún deporte.
- El 48% de los que tienen 30 o más años no practica ningún deporte.

Se elige al azar a una persona de este colectivo. Se pide la probabilidad de que:

a) Sea joven (menos de 30 años) y practique algún deporte.
b) Sea joven y no practique ningún deporte.
c) Tenga 30 o más años y practique algún deporte.
d) Tenga 30 o más años y no practique ningún deporte.
e) Practique algún deporte.
f) No practique ningún deporte.
g) Si la persona elegida practica algún deporte, ¿cuál es la probabilidad de que sea joven?
h) Si la persona elegida no practica ningún deporte, ¿cuál es la probabilidad de que sea joven?

Ejercicio 7. Una empresa compra el 80% de ciertas piezas a un proveedor que le entrega el género con retraso el 10% de les veces. A su vez, por motivos de calidad, la empresa devuelve el 20% de las partidas de este proveedor que llegan con retraso. ¿Cuál es la probabilidad de que una partida de estas piezas haya sido comprada a este proveedor, haya llegado con retraso y se haya tenido que devolver?

Ejercicio 8. Una prueba de selección de personal consta de dos partes: la primera consiste en un test psicotécnico y la segunda de una entrevista personal. Si de 100 persones el 60% supera el test y de estas últimas el 30% supera la entrevista:

a) ¿Cuál es la probabilidad de que una de estas 100 persones, escogida al azar, haya superado les dos pruebas?
b) ¿Cuál es la probabilidad de que no haya superado la entrevista pero sí el test?

Ejercicio 9. En cierta localidad la probabilidad de que una persona compre un diario es 0,4; la probabilidad de que compre una revista 0,2 y la probabilidad de que compre ambos 0,08.

a) ¿Cuál es la probabilidad de que compre alguna de estas 2 publicaciones?

b) Comprar un diario y comprar una revista, ¿son sucesos mutuamente excluyentes?
c) Comprar un diario y comprar una revista, ¿son sucesos independientes?

Ejercicio 10. Con el enunciado del ejercicio 3 determine:

a) ¿Son independientes los sucesos leer el diario B y leer el diario C?
b) Si sabemos que una persona lee el diario A, ¿la probabilidad de que lea B queda modificada?
c) ¿Cuál es la probabilidad de que un lector de B y C lea también A?
d) Sabiendo que un entrevistado lee B, ¿cuál es la probabilidad de que lea A?
e) Sabiendo que un entrevistado lee por lo menos uno de los tres diarios, ¿cuál es la probabilidad de que lea A?

Ejercicio 11. Las probabilidades de que dos personas, que actúan independientemente, lleguen a tiempo a coger el tren se estiman en 0,9 y 0,8, respectivamente. Halle la probabilidad de que:

a) Lleguen ambas.
b) No llegue ninguna.
c) Sólo llegue una.

Ejercicio 12. Un recién graduado que ha solicitado empleo en dos compañías, A y B, estima que tiene doble probabilidad de ser contratado por A que por B y que la probabilidad de que no lo contraten ninguna de las dos es 0,28. Si las decisiones de las dos compañías son independientes, ¿cuál es la probabilidad de que lo contrate A?

Ejercicio 13. Un comercio vende 3 modelos de lector de DVD: V1, V2 y V3. Las probabilidades de venta de cada uno de ellos son: P(V1)=0,3; P(V2)=0,2; y P(V3)=0,5. La probabilidad de avería durante el período de garantía para cada uno de ellos son: 0,10; 0,15 y 0,04 respectivamente. Si un cliente devuelve un aparato averiado en el período de garantía, ¿de qué modelo (V1, V2 o V3) hay más probabilidad de que sea dicho aparato?

Ejercicio 14. Se tienen cuatro cajas idénticas que contienen cuatro bolígrafos cada una. La caja C_1 contiene 4 bolígrafos negros; la caja C_2 contiene 3 negros y 1 rojo; la caja C_3 contiene 2 negros y 2 rojos y la caja C_4 contiene 1 negro y 3 rojos. Se elige una caja al azar y se extrae al azar un bolígrafo.

a) Probabilidad de que el bolígrafo extraído sea negro.
b) Probabilidad de que el bolígrafo extraído sea rojo.

c) Halle las probabilidades *a posteriori* de cada una de las cuatro cajas si se sabe que el bolígrafo extraído es negro.
d) Halle las probabilidades *a posteriori* de cada una de las cuatro cajas si se sabe que el bolígrafo extraído es rojo.

Ejercicio 15. Tres contratistas, A, B y C, licitan por un contrato para construir un polideportivo. Se cree que la probabilidad de obtener el contrato es la misma para cada uno de ellos. La probabilidad de que finalicen las obras en la fecha prevista es 0,85 si el contrato lo consigue A, 0,55 si lo consigue B y 0,30 si lo consigue C. Sabiendo que las obras no han finalizado en la fecha prevista, ¿cuál es la probabilidad de que el contrato lo haya conseguido B?

Ejercicio 16. Respecto al medio de transporte se sabe que el 50% de los alumnos de la Universidad de Barcelona que residen fuera de Barcelona utiliza el tren para ir a clase; mientras que de los que residen en Barcelona no hay ninguno que lo utilice. Sabiendo que el 40% de los alumnos de la UB reside fuera de Barcelona, ¿cuál es la probabilidad de que un alumno elegido al azar entre los que no utilizan el tren resida fuera de Barcelona?

Ejercicio 17. Una entidad financiera ha comprobado que el 5% de los clientes con suficiente saldo se equivocan en la fecha del cheque y el 100% de los que no tienen suficiente saldo cometen el mismo error. Si un 85% de los clientes de esta entidad tienen saldo suficiente, ¿cuál es la probabilidad de que al recibir un cheque con fecha equivocada corresponda a un cliente sin saldo?

Ejercicio 18. Periódicamente una academia controla la asistencia de sus colaboradores eligiendo uno al azar y observando si está en el centro. La academia tiene 4 colaboradores. Uno asiste todos los días, dos de ellos sólo asisten la mitad de los días y el último asiste 2 de cada 5 días. Si en el último control se observa que el colaborador elegido está en la academia, ¿cuál es la probabilidad de que sea el que asiste siempre?

Ejercicio 19. En un laboratorio se sabe por experiencia que 1 de cada 25 frascos de cierto medicamento se deterioran al cabo de un mes, por lo que las existencias se someten a controles mensuales. El test utilizado para determinar si un frasco está deteriorado da positivo el 99% de las veces si el frasco está deteriorado y el 2% de las veces cuando el frasco no está deteriorado. Hallar:

a) Probabilidad de que el test dé positivo.
b) Probabilidad de que el test dé negativo.

c) Si el test resulta positivo, ¿cuál es la probabilidad de que la muestra esté deteriorada?
d) Si el test resulta negativo, ¿cuál es la probabilidad de que la muestra no esté deteriorada?
e) ¿Cuál es la probabilidad de que el test dé el resultado correcto?

Ejercicio 20. Un banco dispone de dos sistemas de alarma A y B que funcionan independientemente entre sí. La alarma A tiene una probabilidad de que funcione correctamente del 80%, mientras que B sólo el 65%. ¿Cuál es la probabilidad de que sólo funcione correctamente una de las dos alarmas?

Ejercicio 21. Una empresa del sector eléctrico tiene un consejo directivo formado por 10 personas, 2 de las cuales son mujeres. Para evaluar las repercusiones que tendría para la empresa la aplicación de un nuevo sistema de tarifas, se designa una comisión de estudio formada por cuatro personas pertenecientes al consejo directivo. ¿Cuál es la probabilidad de que, si se eligen al azar, todos los miembros de dicha comisión sean hombres?

Ejercicio 22. En un multicine, la experiencia indica que el 75% de los asistentes compran palomitas y el 40% compran alguna bebida, siendo la compra de los dos productos independiente. Si sabemos que un espectador ha comprado sólo uno de los productos, ¿cuál es la probabilidad de que haya comprado sólo palomitas?

Ejercicio 23. La probabilidad de que un individuo sea usuario habitual de Internet es 0,3; la probabilidad que sea usuario de la biblioteca municipal es 0,15 y la de que sea usuario habitual de por lo menos uno de estos dos servicios es del 0,4. Se elige al azar un individuo, si es usuario de la biblioteca municipal, ¿cuál es la probabilidad aproximada de que sea usuario de Internet?

Tema 8. VARIABLE ALEATORIA UNIDIMENSIONAL

Variable aleatoria: discreta y continua
Distribución de probabilidad: función de cuantía y función de densidad
Función de distribución
Esperanza matemática y varianza. Variable estandarizada

El conjunto de todos los resultados posibles de un fenómeno aleatorio constituye una población estadística que puede estar formada por resultados cualitativos o cuantitativos. Resulta conveniente asociar estos resultados a valores numéricos para facilitar su representación y análisis.

El concepto de variable aleatoria permite relacionar cada uno de los resultados de un experimento aleatorio con un valor numérico. Con esta transformación se logra caracterizar la población estadística mediante una función o modelo matemático que recoge las descripciones numéricas de los resultados del fenómeno aleatorio junto con sus respectivas probabilidades. Este modelo recibe el nombre de DISTRIBUCIÓN DE PROBABILIDAD de una variable aleatoria.

Las variables aleatorias se clasifican en DISCRETAS y CONTINUAS. La distribución de probabilidad de las discretas se denomina FUNCIÓN DE CUANTÍA P(x) y la de las continuas FUNCIÓN DE DENSIDAD, f(x).

En ambos casos, la FUNCIÓN DE DISTRIBUCIÓN F(x) es la expresión matemática que recoge la probabilidad de que la variable aleatoria tome valores inferiores o iguales a un valor concreto.

La información contenida en las distribuciones de probabilidad de las variables aleatorias puede resumirse en unas pocas medidas que pongan de manifiesto características propias de la distribución. Las más significativas son el valor esperado de X o ESPERANZA MATEMÁTICA**,** y la VARIANZA y DESVIACIÓN ESTÁNDAR poblacional. Conviene destacar la similitud formal y las diferencias conceptuales entre estas medidas poblacionales y las descriptivas.

VARIABLE ALEATORIA

Una variable aleatoria es una función definida entre el espacio muestral y el conjunto de los números reales.

Ejemplo:

Experimento: se lanzan 2 monedas
Variable aleatoria: X= {nº de cruces}

X: E	→	R
(c,c)	→	0
(c,+)	→	1
(+,c)	→	1
(+,+)	→	2

La definición de una variable aleatoria permite asignar un valor numérico a cada uno de los resultados de un experimento aleatorio.

La variable aleatoria puede ser: $\begin{cases} \text{Discreta} \\ \text{Continua} \end{cases}$

DISTRIBUCIÓN DE PROBABILIDAD

DISCRETA: La distribución de probabilidad de un variable discreta recibe el nombre de **Función de Cuantía, P(X)**.

X	P(x)
x_1	$P(x_1)$
x_2	$P(x_2)$
...	...
X_i	$P(x_i)$
...	...
X_n	$P(x_n)$

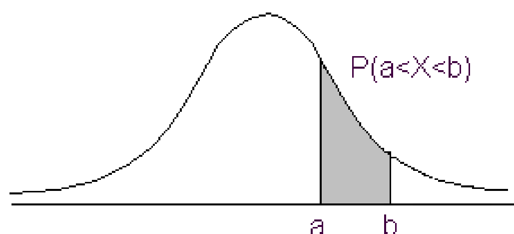
Asocia a cada valor x_i su probabilidad $P(x_i)=P(X=x_i)$.

La función de cuantía siempre verifica:

- $P(x)\geq 0 \ \forall x$. Siempre toma valores no negativos.
- $\sum_{"x} P(x) = 1$. La probabilidad total es igual a 1.

CONTINUA: La distribución de probabilidad de una variable continua queda descrita por una función continua que recibe el nombre de **Función de Densidad, f(x),** tal que:

- $f(x) \geq 0 \ \forall x$. Toma valores no negativos.
- $\int_{-\infty}^{+\infty} f(x)dx = 1$. El área definida por esta función es igual a 1 y, por lo tanto, representa la probabilidad total.



Características:

- $P(x=a) = 0$ La probabilidad de un valor concreto es siempre 0.
- $P(a<X \leq b) = \int_{a}^{b} f(x)\,dx$ La probabilidad de un intervalo es el área definida por la función de densidad.
- $P(a \leq X \leq b) = P(a<X \leq b) = P(a \leq X<b) = P(a<X<b)$

FUNCIÓN DE DISTRIBUCIÓN, F(X)

Recoge la probabilidad acumulada hasta un valor x:

$F(x) = P(X \leq x) = P[-\infty, x]$.

$$F(x) = P(X \leq x) = \begin{cases} \sum_{x \leq xi} P(x_i) & \text{Discreta} \\ \int_{-\infty}^{x} f(x)dx & \text{Continua} \end{cases}$$

Propiedades:

- La función es no negativa. $0 \leq F(x) \leq 1$.
- La función es no decreciente. Dados dos valores $a < b$ entonces $F(a) \leq F(b)$.
- La función converge a cero por la izquierda. $F(-\infty) = 0$.
- La función converge a 1 por la derecha. $F(+\infty) = 1$.
- $P(a < X \leq b) = F(b) - F(a)$.
- $P(X>a) = 1 - F(a)$.

ESPERANZA MATEMÁTICA Y VARIANZA

Permiten resumir la distribución de probabilidad de una variable aleatoria. Reciben el nombre de Parámetros.

Esperanza Matemática o Valor Esperado de X, $E(X)=\mu$

Es el valor medio teórico de la distribución de probabilidad.

Se obtiene:

$$\mu = E(X) = \begin{cases} \sum_{\forall x} x\,P(x) & \text{Discreta} \\ \int_{-\infty}^{+\infty} x\,f(x)\,dx & \text{Continua} \end{cases}$$

Se interpreta como la media de los valores de la variable aleatoria que obtendríamos si el experimento se realizara infinitas veces.

Propiedades:

- La esperanza matemática es el centro de gravedad de la distribución de probabilidad. $E(X-\mu)=0$
- $E(a)=a$. La esperanza matemática de una constante es la constante.
- $E(X+a)=E(X)+a$. La esperanza matemática queda afectada por los cambios de origen.
- $E(bX)=bE(X)$. También le afectan los cambios de escala.
- $E(a+bX)=a+bE(X)$.
- $E(aX+bY)=aE(X)+bE(Y)$ siendo X e Y dos variables aleatorias y, a y b, dos constantes.

Varianza, $V(X)=\sigma^2$

Mide la dispersión de la distribución de probabilidad alrededor de μ.

$$V(X)=\sigma^2=E(X-\mu)^2=\begin{cases} \sum_{\forall x} (x-\mu)^2 P(x) = \sum_{\forall x} x^2 P(x) - \mu^2 & \text{Discreta} \\ \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx = \int_{-\infty}^{+\infty} x^2 f(x)\,dx - \mu^2 & \text{Continua} \end{cases}$$

Propiedades:

- $V(X) \geq 0$
- $V(a) = 0$
- $V(b\,X) = b^2 \cdot V(X)$
- $V(a + b\,X) = b^2 \cdot V(X)$

Desviación estándar, D(x) = σ

$$D(X) = \sigma = +\sqrt{\sigma^2}$$

Presenta las mismas propiedades que la varianza: es no negativa y sólo queda afectada por los cambios de escala.

EJERCICIOS TEMA 8

Ejercicio 1. De un colectivo compuesto por un 60% de hombres se eligen al azar con devolución 3 personas. Establezca la distribución de probabilidad de la variable aleatoria X = "Número de mujeres elegidas".

Ejercicio 2. Un llavero tiene 5 llaves, de las cuales sólo 2 abren una determinada puerta, y se van probando, una a una, hasta conseguir abrir. Se define la variable aleatoria X = 'Número de llaves probadas". Determine la distribución de probabilidad de X bajo los siguientes supuestos:

l) se separan las llaves probadas;
m) las llaves probadas quedan de nuevo mezcladas con las restantes y es imposible diferenciarlas.

Ejercicio 3. Se lanza una moneda hasta que sale cara con un máximo de 5 lanzamientos. Determine la distribución de probabilidad de la variable aleatoria X="Número de lanzamientos realizados".

Ejercicio 4. Un servicio de atención al cliente estima que la distribución de probabilidad de la variable X= "Número de clientes atendidos por hora" es la siguiente:

X_i	0	1	2	3	4	5	6	7
$P(x_i)$	0,01	0,12	0,22	0,32	0,20	0,08	0,03	0,02

a) Halle la función de Distribución de la variable X.
b) ¿Cuál es la probabilidad de que en una hora utilice este servicio algún cliente?
c) ¿Cuál es la probabilidad de que en una hora utilice este servicio un mínimo de cinco clientes?
d) ¿Cuál es la probabilidad de que en una hora sean atendidos más de dos y como máximo cinco clientes?
e) Si se sabe que en una hora determinada algún cliente ha utilizado el servicio, ¿cuál es la probabilidad de que hayan sido más de tres?

Ejercicio 5. Una variable aleatoria X queda caracterizada por la función: $P(x)=k/x$ si x=1, 2, 3 y 4.

a) Determine el valor de k.
b) Obtenga la función de distribución.
c) Calcule $P(1 \leq X<3)$, $P(2<X \leq 4)$, $P(X>2)$, $P(X<2/X<4)$.

Ejercicio 6. La función de cuantía de una variable aleatoria X es:

$$P(x)=\begin{cases} 0{,}05 & \text{para} \quad x=1 \quad y \quad x=8 \\ 0{,}10 & \text{para} \quad x=2 \quad y \quad x=7 \\ 0{,}15 & \text{para} \quad x=3 \quad y \quad x=6 \\ 0{,}20 & \text{para} \quad x=4 \quad y \quad x=5 \\ 0 & \text{Otros casos} \end{cases}$$

a) Compruebe que es función de cuantía.
b) Determine la función de distribución.
c) Obtenga las siguientes probabilidades: $P(X<4)$, $P(1<X\leq 4)$, $P(X>2)$, $P(3\leq X<6)$, $P(X\geq 3/X<7)$.

Ejercicio 7. Se sabe que en promedio uno de cada 10 clientes que entran en cierto establecimiento realiza una compra. La variable X = 'Nº de clientes que entran hasta que compra uno (incluido éste)' tiene la siguiente función de distribución:

$$F(x) = 1-0{,}90^{x} \quad \text{si} \quad x = 1, 2, 3, ...$$

Se pide:

a) Probabilidad de que el décimo cliente que entre sea el primero que compre.
b) Probabilidad de que hayan entrado como máximo 4 clientes hasta que compra uno.
c) Probabilidad de que hayan entrado por lo menos 3 clientes hasta que compra uno.
d) Si se sabe que ya han entrado más de 4 clientes, ¿cuál es la probabilidad de que entren como máximo 7 considerando el que compra?

Ejercicio 8. La función de distribución de una variable aleatoria es:

$$F(x)=\begin{cases} 0 & x<1/8 \\ 0{,}2 & 1/8 \leq x<1/4 \\ 0{,}9 & 1/4 \leq x<3/8 \\ 1 & x \geq 3/8 \end{cases}$$

a) Indique si la variable es continua o discreta.
b) Obtenga la función de probabilidad.
c) Represente gráficamente las dos funciones.

Ejercicio 9. Sea la función,

$$f(x)=\begin{cases} 3x^2/2 & -1 \leq x \leq +1 \\ 0 & \text{Otros casos} \end{cases}$$

a) Compruebe si es función de densidad.
b) Calcule la probabilidad de que X sea superior a -0,5 e inferior a 0,25.
c) Obtenga la función de distribución.

Ejercicio 10. La función de densidad que caracteriza a la variable aleatoria X es:

$$f(x)=\begin{cases} 1/k & 1 \le x \le 5 \\ 0 & \text{Otros casos} \end{cases}$$

a) Determine el valor de k.
b) Obtenga la función de distribución.
c) Calcule las siguientes probabilidades: $P(X \le 3)$, $P(1{,}5 \le X < 4)$, $P(X>2)$, $P(X>3/X>2)$.

Ejercicio 11. Sea la función,

$$f(x)=\begin{cases} x^2/9 & 0 \le x \le k \\ 0 & \text{Otros casos} \end{cases}$$

a) Determine k para que la función sea función de densidad.
b) Calcule la probabilidad de que X sea superior a 0,5 sabiendo que es inferior a 2,5.

Ejercicio 12. Dadas las siguientes funciones de distribución:

A) $$F(x)=\begin{cases} 0 & x<0 \\ 0{,}1 & 0 \le x<2 \\ 0{,}3 & 2 \le x<3 \\ 0{,}6 & 3 \le x<4 \\ 1 & x \ge 4 \end{cases}$$

B) $$F(x)=\begin{cases} 0 & x<0 \\ \dfrac{x^3}{64} & 0 \le x \le 4 \\ 1 & x>4 \end{cases}$$

a) Indique, en cada caso, si la variable es continua o discreta.
b) Calcule, para estas variables, $P(X \le 2)$, $P(1 \le X<3)$, $P(X>2)$, $P(X>1/X<2{,}5)$.

Ejercicio 13. En determinada estación de servicio la variable aleatoria X='Demanda semanal de gasolina en miles de litros' tiene la siguiente función de distribución:

$$F(x)=\begin{cases} 0 & x<0 \\ 1{,}2x-0{,}2x^2 & 0 \le x \le 1 \\ 1 & x>1 \end{cases}$$

Determine:

a) La probabilidad de que la demanda sea inferior a 0,5 (miles de litros).
b) La probabilidad de que la demanda sea inferior a 0,5 y superior a 0,3.
c) La probabilidad de que la demanda sea exactamente 0,2.
d) La probabilidad de que la demanda sea superior a 0,3 sabiendo que es inferior a 0,5.
e) La probabilidad de que la demanda supere los 0,4 (miles de litros).

Ejercicio 14. En los paquetes de harina de 10 kg se comete un error aleatorio en el peso, X (en kg), cuyo comportamiento queda recogido por la función de distribución:

$$F(x)=\begin{cases} 0 & x<-1 \\ \dfrac{x^3}{2}+k & -1 \leq x \leq 1 \\ 1 & x>1 \end{cases}$$

a) Determine el valor de k.
b) Calcule la probabilidad de que un paquete de harina supere 10,5 kg.
c) En una partida de 1000 paquetes ¿qué proporción de paquetes se espera que pesen más de 9 kg y cuarto?
d) Sabiendo que un paquete supera los 9 kg y medio, ¿cuál es la probabilidad de que no alcance los 10 kg y cuarto?
e) ¿Cuál es el peso mínimo de un paquete para poder decir que está entre el 30% de los que más pesan?

Ejercicio 15. Un taller estima que la distribución de probabilidad de la variable X= "Tiempo empleado en el empaquetado de un determinado tipo de piezas" es la siguiente:

$$f(x)=\begin{cases} 6x(1-x) & 0 \leq x \leq 1 \\ 0 & \text{Otros casos} \end{cases}$$

a) ¿Cuál es el tiempo medio del empaquetado?
b) ¿Considera que este valor esperado es una buena medida de síntesis del comportamiento de X?
c) Se reajusta el proceso y el nuevo tiempo del empaquetado es Y= 1,5X-0,5 manteniendo la misma distribución de probabilidad. Calcule el nuevo valor esperado del tiempo.
d) ¿El ajuste introducido ha conseguido mejorar la regularidad del proceso?

Ejercicio 16. Una variable aleatoria discreta, X, tiene la siguiente función de distribución:

$$F(x)=\begin{cases} 0 & x<3 \\ 1/12 & 3 \leq x<5 \\ 1/3 & 5 \leq x<7 \\ 3/4 & 7 \leq x<8 \\ 1 & x \geq 8 \end{cases}$$

a) Calcule $P(3,5 < X \leq 5)$.
b) Calcule la probabilidad de que X sea inferior a 7 sabiendo que ha tomado un valor superior a 4.
c) Determine la función de cuantía.
d) Calcule la mediana, el valor esperado y la desviación estándar de X.

Ejercicio 17. Un inversor está considerando tres alternativas para invertir 1000€. Se estiman los siguientes resultados:

Alternativa 1: un beneficio de 10.000€. con probabilidad de 0,15 y una pérdida de 1.000€ con probabilidad de 0,85
Alternativa 2: un beneficio de 1.000€. con probabilidad de 0,5 y una pérdida de 500€ con probabilidad de 0,5
Alternativa 3: un beneficio seguro de 400€

a) ¿En cuál de las tres opciones su beneficio esperado es mayor?
b) ¿Aconsejaría sin dudar al inversor que escogiera esta opción?

Ejercicio 18. El gerente de una fábrica está considerando cambiar una máquina cuyo número de averías semanales presenta la siguiente distribución de probabilidad.

Nº de averías	0	1	2	3	4
$P(x_i)$	0,1	0,26	0,42	0,16	0,06

La decisión será cambiarla si el coste esperado de las reparaciones semanales supera los 200 Euros. ¿Qué decisión tomará si la reparación de cada avería cuesta 150 Euros?

Ejercicio 19. Una compañía fabrica clips y los comercializa en paquetes de aproximadamente 50 unidades. El número de clips por paquete varía según la siguiente distribución de probabilidad:

Nº de clips	47	48	49	50	51	52
Probabilidad	0,05	0,10	0,25	0,30	0,20	0,10

a) ¿Cuál es la probabilidad de que un paquete contenga más de 48 clips y como máximo 51?
b) Sabiendo que un paquete contiene menos de 50 clips, ¿cuál es la probabilidad de que contenga más de 48?
c) ¿Cuál es el número esperado de clips por paquete y su varianza?
d) Si el coste de un paquete es 15+2X, donde X es el nº de clips, ¿cuál es el coste esperado y la varianza del coste de estos paquetes?

Ejercicio 20. Calcule el valor esperado, la varianza y la desviación estándar de las variables aleatorias de los ejercicios 9, 11 y 12.

Ejercicio 21. X es una variable aleatoria que toma los valores: -2, 1, 2 y 4 con probabilidades: (2k-3)/10, (k-2)/10, (k-1)/10 y (k+1)/10, respectivamente.

a) Se pide:
b) Valor de k.
c) Valor esperado, mediana y moda de X.
d) Varianza y desviación estándar de X.
e) Distribución estandarizada y comprobar que su valor esperado es 0 y la varianza 1.

Ejercicio 22. Una lotería con 100 boletos da 1 premio de 500 €, 2 premios de 100 € y 6 premios de 50 €. Si cada boleto cuesta 10 €, ¿cuál es el valor esperado de este juego por boleto?

Ejercicio 23. Un jugador puede lanzar 2 veces una moneda equilibrada. Gana 3 € si salen 2 caras y 1 € si sólo sale una cara. Para que el juego sea justo ¿cuánto tiene que perder si no salen caras?

Ejercicio 24. La función de densidad de la variable aleatoria X ="Importe semanal facturado en miles de €" en un establecimiento es:

$$f(x)=\begin{cases} \dfrac{1+x}{12} & x \in [0,4] \\ 0 & x \notin [0,4] \end{cases}$$

a) Calcule el importe que en promedio se espera facturar por semana.
b) ¿Qué importe máximo espera conseguir el 40% de las semanas con menos facturación?
c) ¿Cuál será, aproximadamente, el mínimo importe que facturará el 25% de las semanas con mayor facturación?
d) ¿Cuál es la varianza y la desviación estándar de X?

e) Se estima que la facturación se ha transformado y ha pasado a ser $0{,}75X+0{,}25$. ¿Es más regular la facturación que se espera obtener en esta nueva situación?

Tema 9. DISTRIBUCIONES BINOMIAL Y NORMAL

Distribución Binomial
Distribución Normal.

Algunas distribuciones de probabilidad modelizan un gran número de fenómenos aleatorios en el ámbito social, económico, biológico etc..,por ello reciben un nombre propio.

En este tema se incluye una distribución de probabilidad de tipo discreto: la distribución BINOMIAL y una de tipo continuo, la distribución NORMAL.

Para cada una de estas distribuciones se estudiará la función de cuantía o de densidad, la función de distribución, los parámetros que las definen, así como el cálculo de probabilidades y utilización de las tablas.

DISTRIBUCIÓN DICOTÓMICA O DE BERNUILLI

A menudo estamos interesados en estudiar el comportamiento de alguna variable que sólo puede tomar dos posibles resultados, por ejemplo, al observar un producto comprobar si cumple o no unas determinadas condiciones de calidad, al observar un paciente ver si presenta o no una determinada enfermedad, al lanzar una moneda ver si el resultado es cara o cruz. En estas situaciones los sucesos posibles son dos alternativas complementarias o dicotómicas y su distribución de probabilidad se denomina distribución de Bernoulli.

Si consideramos un experimento aleatorio con sólo dos posibles resultados, 'éxito' o 'no éxito', y la probabilidad de éxito es p y la probabilidad de no éxito es q=1-p, la variable:

X="Número de éxitos obtenidos en una realización del experimento" presenta una distribución Dicotómica con función de cuantía:

$$P(x) = p^x \, q^{1-x} \qquad \text{para } x = 0, 1$$

X	P(X)
0	1-p=q
1	p

Características:

- Para identificar una distribución dicotómica concreta dentro de la familia de distribuciones dicotómicas basta conocer el valor del parámetro p, probabilidad de éxito.
- La distribución de X se abrevia X~D(p)
- El valor esperado de X es E(X)=p
- La varianza de X es V(X)=pq

DISTRIBUCIÓN BINOMIAL

La distribución Binomial se obtiene como generalización del proceso de Bernoulli. Por ejemplo, supongamos que se lanza una moneda 10 veces y que se define la variable aleatoria X como el 'número de veces que ha salido cara en los 10 lanzamientos'. En este caso, la variable aleatoria puede tomar los valores enteros de 0 a 10.

Si suponemos que se realizan n lanzamientos independientes y que la probabilidad de obtener cara es un valor p que se mantiene constante en todos los lanzamientos, entonces la variable aleatoria X, que recoge el número total de caras obtenidas, presenta una distribución Binomial de parámetros n y p.

Por tanto, consideremos un experimento aleatorio tal que:

- Cada vez que se realiza el experimento ocurre uno y sólo uno de los siguientes resultados: 'éxito' o 'no éxito'.
- Cada vez que se repite el experimento el resultado es independiente del obtenido en las realizaciones anteriores, por lo que la probabilidad de éxito, p, es constante en cada prueba.
- Se realiza el experimento n veces.

Se define la variable:

X = "Número de éxitos obtenidos en las n realizaciones del experimento"

La variable X presenta distribución Binomial de parámetros n y p, y su función de cuantía es:

$$P(x) = \binom{n}{x} p^x q^{n-x} \quad \text{para } x = 0, 1, 2, ..., n \text{ y } q = 1-p$$

Características:

- Para identificar una distribución binomial concreta dentro de la familia de distribuciones binomiales basta con conocer los valores de los parámetros, n y p, que la caracterizan.
- La distribución de X se abrevia X~B(n; p).
- La distribución dicotómica es un caso particular de distribución binomial con parámetros n=1 y p: X~B(1; p)=D(p).
- La distribución B(n;p) se obtiene como suma de n distribuciones dicotómicas independientes de parámetro p.
- El valor esperado de de X es E(X)=n p.
- La varianza de X es V(X) = npq.
- La distribución binomial es reproductiva en el parámetro p: si se suman dos o más variables binomiales independientes con el mismo parámetro p, la variable resultante tiene distribución: $B(\sum_{i=1}^{k} n_i, p)$.
- Para cualquier valor de n, la distribución de X es simétrica si p=0,5; presenta asimetría positiva si p < 0,5 y asimetría negativa si p > 0,5. La asimetría se reduce a medida que p se aproxima a 0,5. Asimismo, para cualquier valor de p la asimetría disminuye cuando aumenta el valor de n.

DISTRIBUCIÓN NORMAL

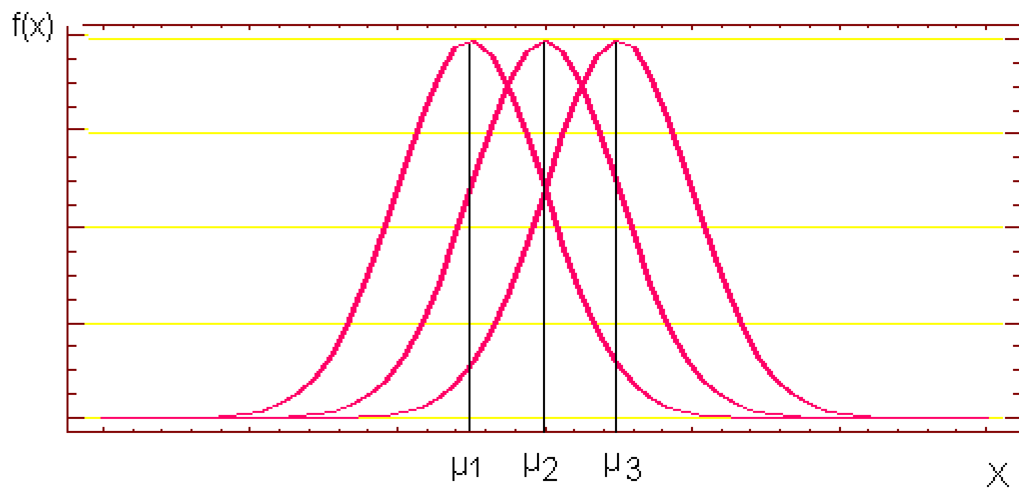
Es la distribución de probabilidad continua más importante ya que recoge el comportamiento poblacional de gran número de variables. Además la distribución Normal es la base de la inferencia estadística ya que la distribución de probabilidad de la mayoría de los estadísticos muestrales converge en esta distribución cuando el tamaño de la muestra es suficientemente elevado.

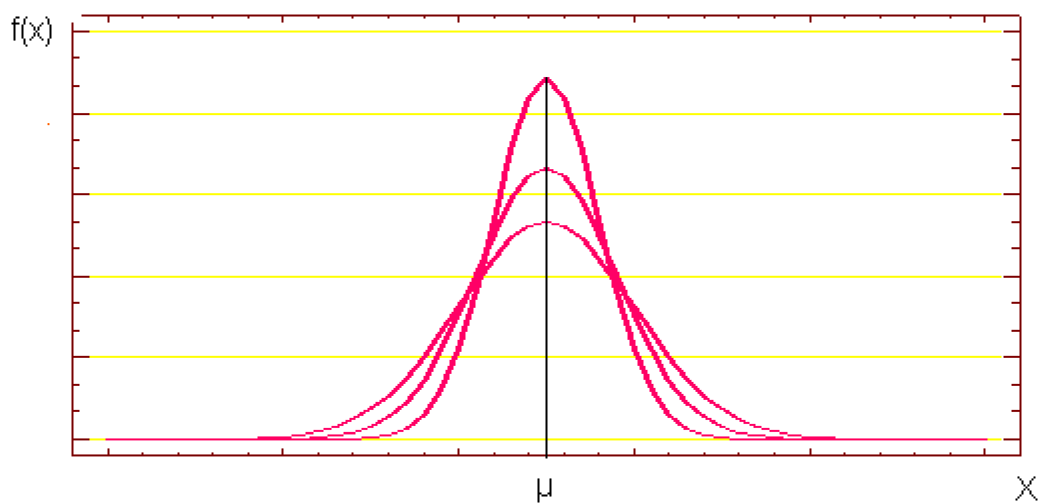
Una variable aleatoria continua X presenta una distribución normal de parámetros μ y σ si su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2} \quad \forall x \in R$$

donde: $\mu \in R$, $\sigma > 0$, п=3,14 y e= 2,71

Existe una familia de infinitas distribuciones normales:





Características:

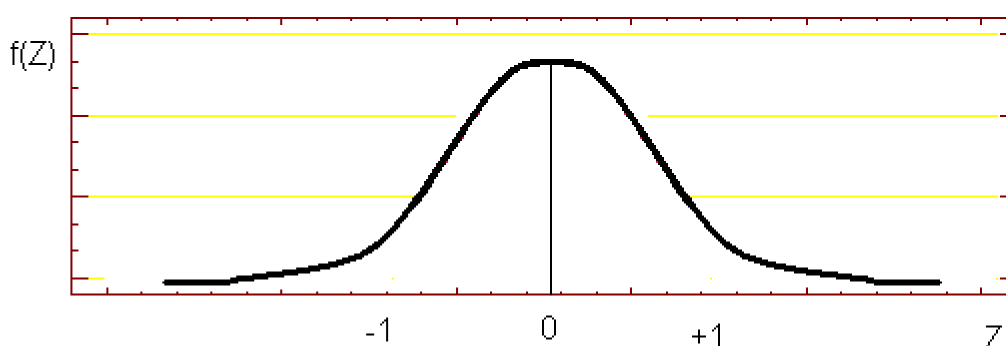
- La distribución normal queda identificada por dos parámetros: su valor esperado, μ y su desviación estándar, σ.
- La distribución de X se abrevia: $X \sim N(\mu, \sigma)$.
- La variable X puede tomar cualquier valor Real de $-\infty$ a $+\infty$.
- La distribución de X es campanoide y simétrica:
- El coeficiente de asimetría es 0.
- Esperanza matemática, mediana y moda coinciden.
- $P(X<\mu)=P(X>\mu)=0,5$.
- La distribución de X es mesocúrtica y su coeficiente de curtosis es 0.
- La distribución normal presenta dos puntos de inflexión en $\mu-\sigma$ y $\mu+\sigma$.
- Es asintótica respecte al eje de abscisas.

- La distribución normal es reproductiva; al sumar o restar dos o más variables normales independientes se obtiene una nueva variable normal de parámetros $N(\Sigma\mu, \sqrt{\Sigma\sigma^2})$.

DISTRIBUCIÓN NORMAL ESTANDARIZADA O TIPIFICADA

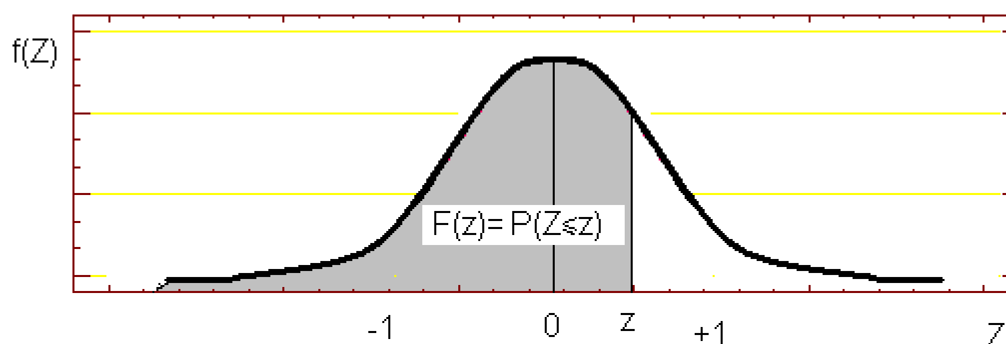
De entre las infinitas curvas normales la correspondiente a los parámetros $\mu=0$ y $\sigma=1$ recibe el nombre de distribución normal estandarizada o tipificada y presenta una especial importancia. Se simboliza $Z \sim N(0, 1)$.

Su función de densidad es $f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2} \quad \forall z\in R$.



Características:

- Es simétrica: $P(Z<0) = P(Z>0) = 0,5$.
- Presenta un máximo en $z=0$.
- Tiene dos puntos de inflexión -1 y +1.
- Cualquier otra variable normal X de parámetros μ y σ se puede transformar en una normal estandarizada simplemente mediante la transformación lineal $Z=\frac{X-\mu}{\sigma}$.
- La función de distribución de esta variable, F(z), está tabulada.



La tabla permite obtener probabilidades de sucesos referidos a cualquier variable normal X~N(μ, σ), ya que:

$$P(a \le X \le b) = P(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma})$$

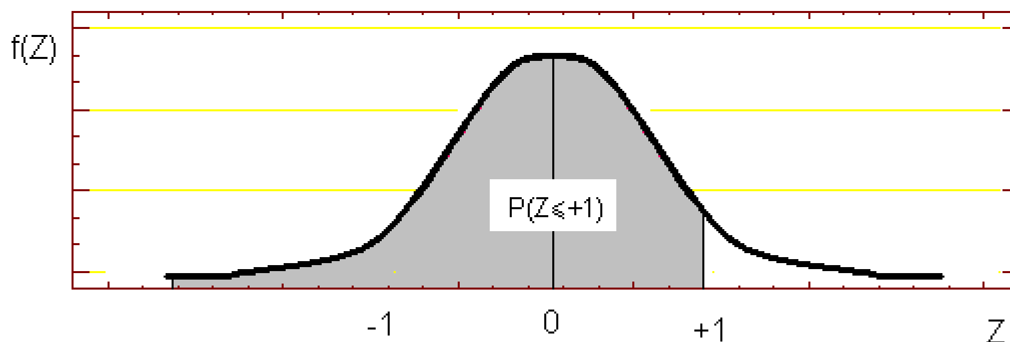
Utilización de la tabla de la distribución Normal

La tabla contiene los valores de la función de distribución de la variable Z~N(0, 1) o normal tipificada para valores positivos de la variable.

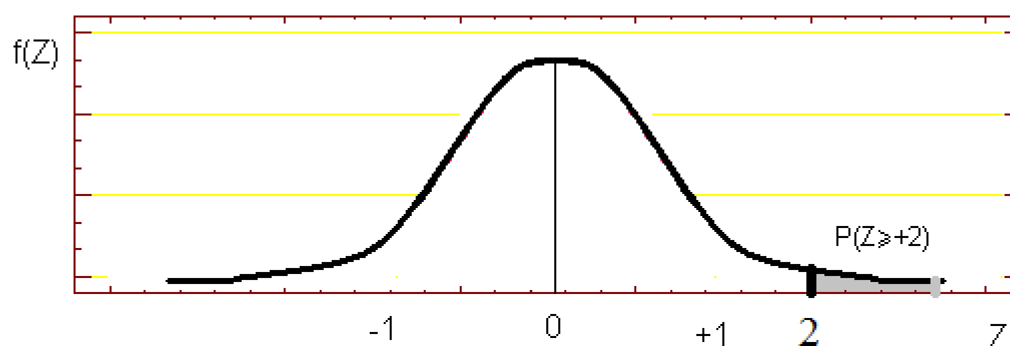
Para cualquier otra variable normal X~N(μ, σ) es necesario transformar X en Z: $Z = \frac{X-\mu}{\sigma}$

Ejemplo: Dada una variable aleatoria X~N(40, σ=10), halle las siguientes probabilidades:

1. $\mathbf{P(X \le 50) = P(Z \le \frac{50-40}{10}) = P(Z \le 1) = F(1)} = 0{,}8413$

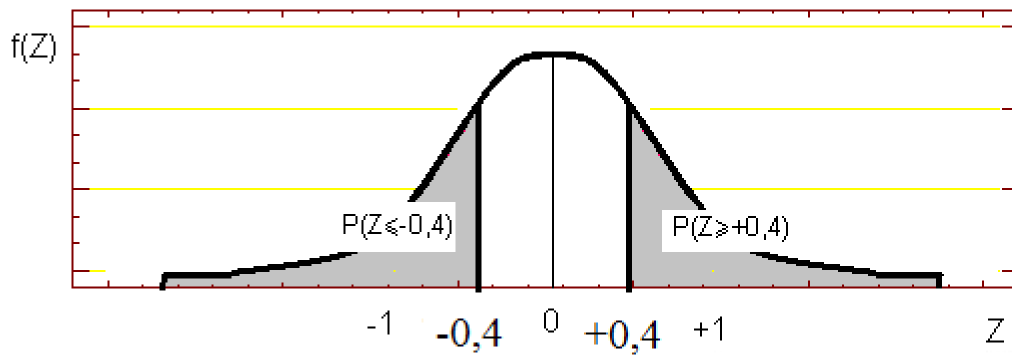


2. $\mathbf{P(X \ge 60) = P(Z \ge \frac{60-40}{10}) = P(Z \ge 2)} = 1 - \mathbf{P(Z \le 2)} = 1 - \mathbf{F(2)} = 1 - 0{,}9772 = 0{,}0228$

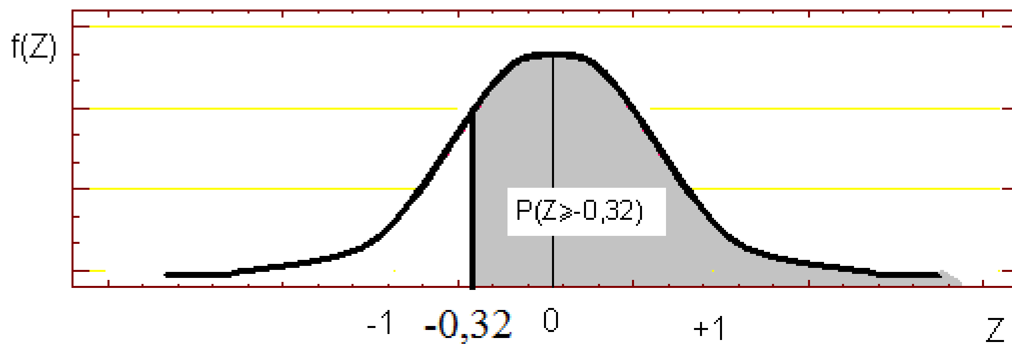


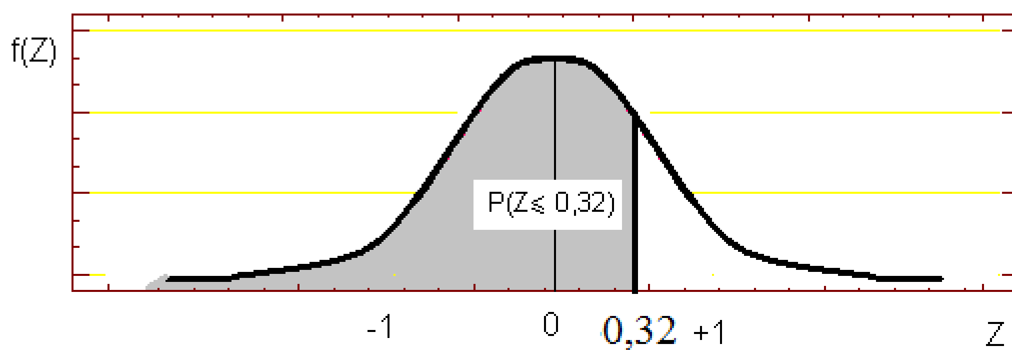
3. $\mathbf{P(X \le 36) = P(Z \le \frac{36-40}{10})} = \mathbf{P(Z \le -0,4)} = \mathbf{P(Z > 0,4)} = 1-\mathbf{P(Z \le 0,4)} = 1-\mathbf{F(0,4)} =$

$= 1-0,6554 = 0,3446$



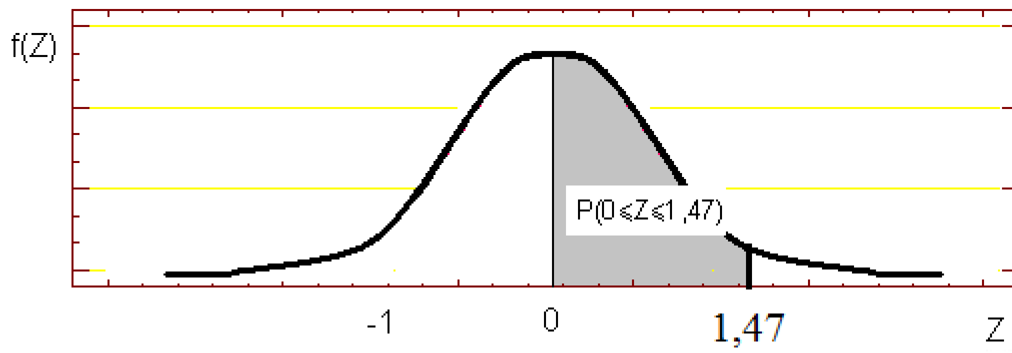
4. $\mathbf{P(X \ge 36,8) = P(Z \ge \frac{36,8-40}{10})} = \mathbf{P(Z \ge -0,32)} = \mathbf{P(Z \le 0,32)} = \mathbf{F(0,32)} = 0,6255$



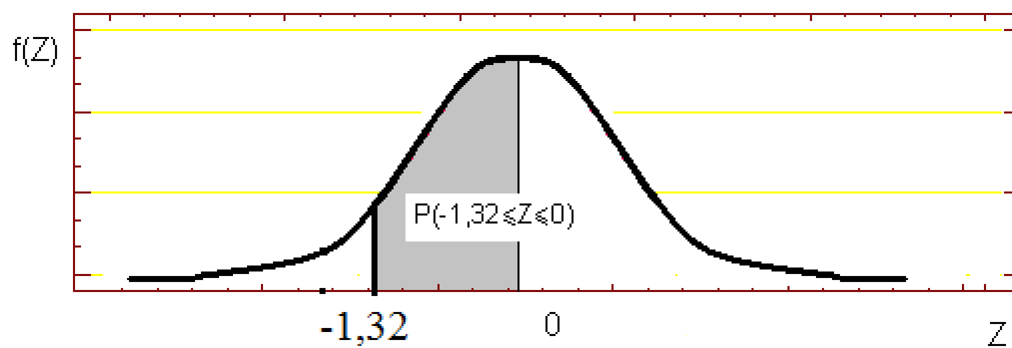


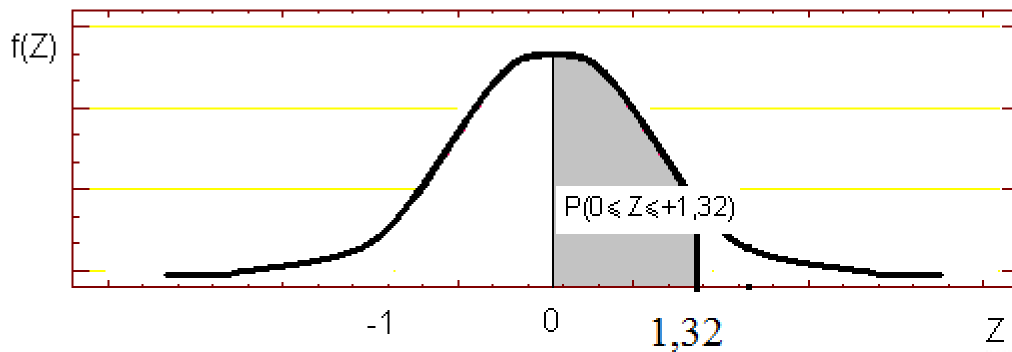
5. $\mathbf{P(40 \le X \le 54,7) = P(\frac{40-40}{10} \le Z \le \frac{54,7-40}{10})} = \mathbf{P(0 \le Z \le 1,47)} =$

$= \mathbf{F(1,47)} - \mathbf{F(0)} = 0,9292 - 0,5 = 0,4292$

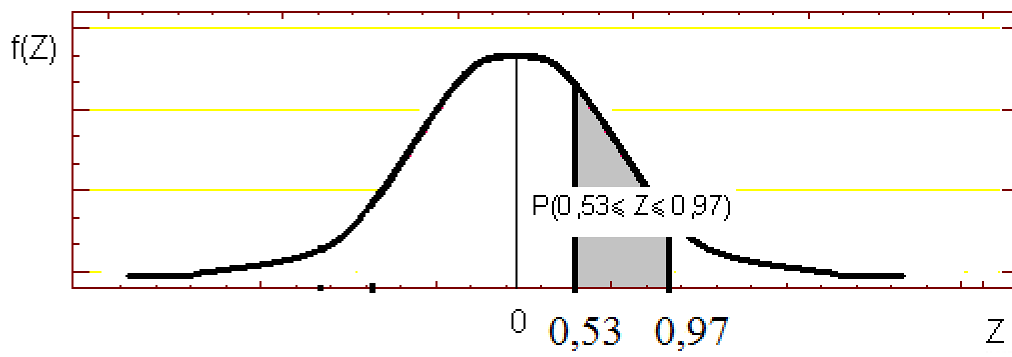


6. $\mathbf{P}(26{,}8 \le \mathbf{X} \le 40) = \mathbf{P}(\frac{26{,}8-40}{10} \le \mathbf{Z} \le \frac{40-40}{10}) = \mathbf{P}(-1{,}32 \le \mathbf{Z} \le 0) = \mathbf{P}(0 \le \mathbf{Z} \le 1{,}32) =$
$= \mathbf{F}(1{,}32) - \mathbf{F}(0) = 0{,}9065 - 0{,}5 = 0{,}4065$



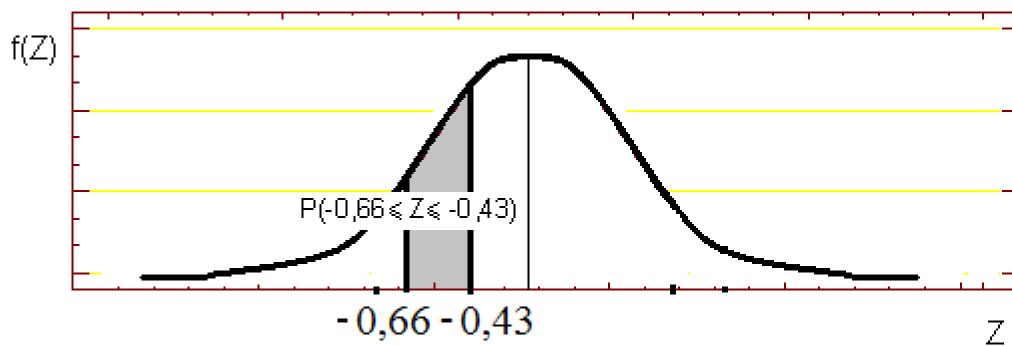


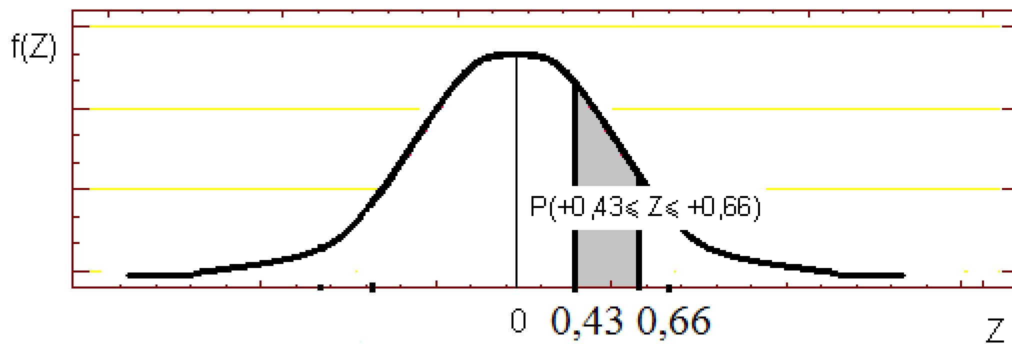
7. $\mathbf{P}(45{,}3 \le \mathbf{X} \le 49{,}7) = \mathbf{P}(\frac{45{,}3-40}{10} \le \mathbf{Z} \le \frac{49{,}7-40}{10}) = \mathbf{P}(0{,}53 \le \mathbf{Z} \le 0{,}97) =$
$= \mathbf{F}(0{,}97) - \mathbf{F}(0{,}53) = 0{,}8339 - 0{,}7019 = 0{,}132$



8. $\mathbf{P}(33,4 \le \mathbf{X} \le 35,7) = \mathbf{P}(\frac{33,4-40}{10} \le \mathbf{Z} \le \frac{35,7-40}{10}) = \mathbf{P}(-0,66 \le \mathbf{Z} \le -0,43) =$

$= \mathbf{P}(0,43 \le \mathbf{Z} \le 0,66) = \mathbf{F}(0,66) - \mathbf{F}(0,43) = 0,7453 - 0,6664 = 0,0789$





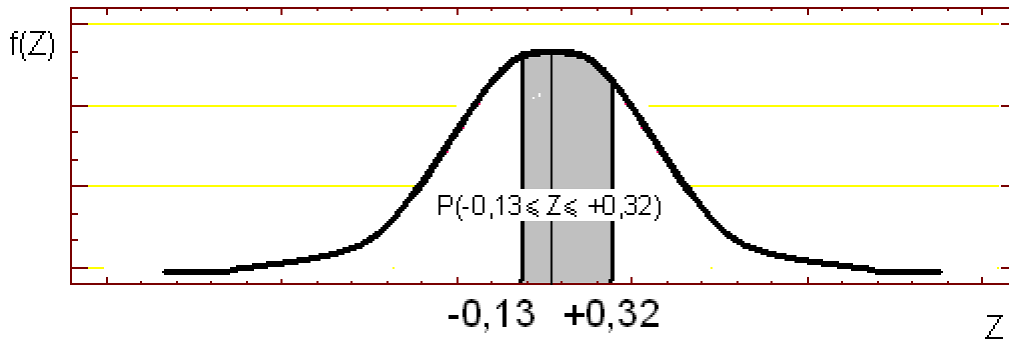
9.

$$\mathbf{P}(38,7 \le \mathbf{X} \le 43,2) = \mathbf{P}\left(\frac{38,7-40}{10} \le \mathbf{Z} \le \frac{43,2-40}{10}\right) = \mathbf{P}(-0,13 \le \mathbf{Z} \le 0,32) =$$

$$= \mathbf{P}(\mathbf{Z} \le 0,32) - \mathbf{P}(\mathbf{Z} \le -0,13) = \mathbf{P}(\mathbf{Z} \le 0,32) - \mathbf{P}(\mathbf{Z} \ge 0,13) =$$

$$= \mathbf{P}(\mathbf{Z} \le 0,32) - [1 - \mathbf{P}(\mathbf{Z} \le 0,13)] = \mathbf{F}(0,32) - 1 + \mathbf{F}(0,13) =$$

$$= 0,6255 - 1 + 0,5517 = 0,1772$$



EJERCICIOS TEMA 9

Ejercicio 1. Obtenga las siguientes probabilidades indicando, en cada caso, la definición y la distribución de la variable:

a) Probabilidad de que exactamente 2 clientes, entre 6 elegidos al azar, paguen con tarjeta de crédito si en general el 40% pagan utilizando este sistema.
b) Probabilidad de que obtengamos como máximo 2 caras al lanzar 5 veces una moneda.
c) Probabilidad de que una familia con 3 hijos elegida al azar tenga 2 o más niñas si consideramos que la probabilidad de tener niño o niña es la misma.
d) Probabilidad de que en un día elegido al azar más de una máquina se estropee si en el taller hay 8 máquinas idénticas con comportamientos independientes y presentan una probabilidad de estropearse de 0,25.
e) Probabilidad de obtener como mínimo 4 resultados impares en 5 lanzamientos de un dado.
f) Probabilidad de obtener 3 o más facturas impagadas si se extraen con reposición 10 de un archivador que contiene 600 facturas pagadas y 200 impagadas.

Ejercicio 2. La tasa de paro de un determinado país es del 10% de la población activa. Si se seleccionan al azar 7 personas determinar la probabilidad de que:

n) Exactamente 2 estén en paro.
o) Como máximo 2.
p) Como mínimo 2.

Ejercicio 3. El departamento de control de calidad sabe que la producción de cierto artículo presenta una proporción de defectuosos del 5%. Si estos artículos se venden en cajas de 10 unidades, calcule la probabilidad de que una caja contenga:

a) Ninguna unidad defectuosa.
b) Entre 2 y 4 unidades defectuosas.
c) Como máximo 1 unidad defectuosa.
d) Como mínimo un 80% de unidades de defectuosas.
e) Como máximo 1 sabiendo que contiene por lo menos 1 unidad defectuosa.
f) Ninguna defectuosa sabiendo que como máximo contiene 3 defectuosas.
g) ¿Cuál es el número más probable de unidades defectuosas por caja?
h) Si un pedido consta de 5 cajas, ¿cuál es la probabilidad de que exactamente el 80% de las cajas no contenga ninguna unidad defectuosa?

Ejercicio 4. Un test consta de 20 preguntas con 4 respuestas cada una de las cuales sólo una es correcta. A cada pregunta correcta se le asigna 1 punto y se restan 0,25 puntos por respuesta incorrecta. Si un estudiante responde a todas las preguntas al azar:

a) ¿Cuál es el número esperado de respuestas correctas y su desviación típica?
b) ¿Cuál es la puntuación esperada y su desviación típica?

Ejercicio 5. De una lista de 1000 números de teléfono se tiene la siguiente información: 350 corresponden a móviles y 250 corresponden al distrito de Gràcia. Se extraen al azar con reposición 5 números de esta lista. Halle la probabilidad de que:

a) Ninguno corresponda a un teléfono móvil.
b) Ninguno corresponda al distrito de Gracia.
c) Ninguno corresponda ni a móvil ni al distrito de Gracia (suponga independencia entre clase de teléfono y distrito).

Ejercicio 6. Una empresa que compra determinados componentes en lotes grandes sólo los acepta si contiene como máximo un 10% de defectuosos. A estos efectos se inspeccionan al azar 20 unidades de cada lote, y éste se acepta si como máximo la proporción de unidades defectuosas es del 10%. Si el proveedor estima que la proporción real de componentes defectuosas es del 5%:

a) ¿Cuál es la probabilidad de rechazar un lote?
b) ¿Cuál es la varianza y la desviación estándar de la variable 'número de piezas defectuosas en la muestra'?
c) Si el pedido consta de 10 lotes, ¿cuál es la probabilidad de rechazar como máximo 1?
d) Si el número de piezas inspeccionadas fuera de 10, ¿varía la probabilidad de rechazar el lote?

Ejercicio 7. Un vendedor ha comprobado que la distribución de probabilidad de sus ventas semanales es:

X	1	2	3	4	5	6
P(X)	0,25	0,35	0,15	0,10	0,05	0,10

Siendo X=nº de coches vendidos.

La empresa le paga una prima semanal de 500 Euros si consigue vender más de 2 coches. Si las ventas semanales son independientes, calcule la probabilidad de que en 1 mes (4 semanas),

a) Consiga como máximo 500 Euros de prima,
b) Consiga más de 1000 Euros de prima.
c) ¿Cuál es el valor esperado y la desviación estándar de la prima mensual que puede obtener este vendedor?

Ejercicio 8. Una compañía tiene dos proveedores de un determinado artículo. El 70% de los pedidos procede del proveedor cuyos envíos suelen contener un 10% de artículos defectuosos y el resto proceden del otro proveedor cuyos envíos suelen contener un 20% de defectuosos. Se recibe un pedido, pero se desconoce su procedencia, se analiza una muestra de 20 artículos de este pedido y se encuentra 1 defectuoso. ¿Cuál es la probabilidad de que el pedido proceda del proveedor con menor tasa de artículos defectuosos?

Ejercicio 9. Dada una variable Normal de parámetros $\mu=10$ $\sigma=2$ obtenga:

a) $P(X\leq 14)$
b) $P(X>11{,}75)$
c) $P(X< 9)$
d) $P(5<X\leq 12)$
e) $P(6\leq X\leq 14)$
f) $P(X>11/X>9{,}5)$

Ejercicio 10. Determine el valor a que verifique:

a) $P(X<a)= 0{,}846$ siendo $X \sim N(50; 15)$
b) $P(X>a)= 0{,}42$ siendo $X \sim N(10; 5)$
c) $P(X<a)= 0{,}25$ siendo $X \sim N(250; 50)$
d) $P(X>a)= 0{,}95$ siendo $X \sim N(20; 8)$

Ejercicio 11. Por experiencia se sabe que la puntuación obtenida en cierto examen es una variable aleatoria Normal con esperanza matemática 11,5 puntos y desviación estándar 5.

a) Si la puntuación mínima para aprobar se fija en 10 puntos ¿qué porcentaje de alumnos aprobará?
b) Si se quiere aprobar al 69,5% de los presentados ¿qué nota mínima se debe exigir?
c) Si las puntuaciones mínima y máxima para obtener un notable son 15,5 y 18, respectivamente ¿cuántos notables se espera contabilizar en una prueba con 200 presentados?

d) ¿Qué puntuación mínima debe tener un examen para estar entre los 5 mejores en una prueba de 500 presentados?

Ejercicio 12. El peso neto (en gr.) de los paquetes de galletas de la marca A se puede modelizar mediante la variable aleatoria X~N(500; 50). Se pide:

a) ¿Qué porcentaje de paquetes supera los 550 gr?
b) ¿Cuál es la probabilidad de que un paquete pese entre 480 y 520 gr?
c) En una partida de 2000 paquetes ¿cuántos se espera que no alcancen los 475 gr?
d) Se considera que un paquete es apto para la venta si pesa más de 440 gr. ¿Cuál es la probabilidad de que un paquete sea apto para la venta?
e) Bajo el supuesto que sólo un 15% de los paquetes se va a considerar **No** aptos para la venta ¿qué peso debe presentar un paquete para considerarlo no apto?
f) Los pedidos se sirven al minorista en cajas de 10 paquetes. ¿Cuál es la probabilidad de que una caja contenga por lo menos 8 paquetes aptos para la venta, es decir con peso superior a 440 gr?

Ejercicio 13. En una estación de servicio se sabe que la demanda mensual de gasolina (en litros) puede modelizarse mediante una Normal de media 150000 litros y desviación típica 10000 litros. Determine la cantidad que hay que tener disponible cada mes si se desea que la probabilidad de satisfacer la demanda sea igual a 0,95

Ejercicio 14. Un mecanógrafo sabe que el tiempo que tarda en mecanografiar una página de un manuscrito es una variable aleatoria aproximadamente normal de media 6 minutos y varianza 1,44.

a) ¿Cuál es la probabilidad de que necesite más de 8 minutos para mecanografiar una página?
b) De un manuscrito de 100 páginas, ¿qué tiempo como máximo necesitará para mecanografiar una de las 20 páginas más cortas?

Ejercicio 15. Las retribuciones del personal de una empresa siguen una distribución normal. Se sabe que el 2% son superiores a 42000 Euros y el 10% inferiores a 15000 Euros. ¿Qué proporción son inferiores a 30000 Euros?

SOLUCIÓN EJERCICIOS

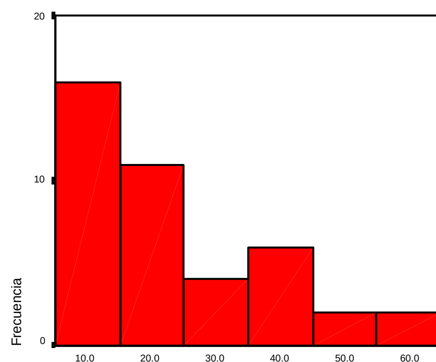
TEMA 1 Y 2

1.

a) 100
b) 23%
c) 18
d) 9 y 1
e) 5 y 6
f) 4
g) 16%
h) 6
i) 4

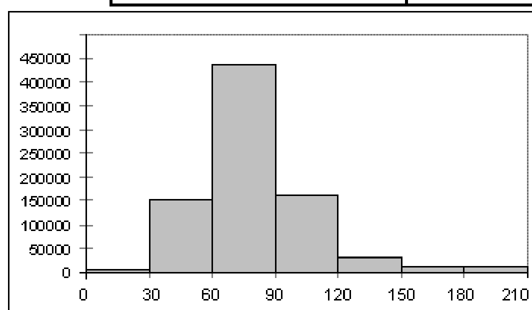
2.

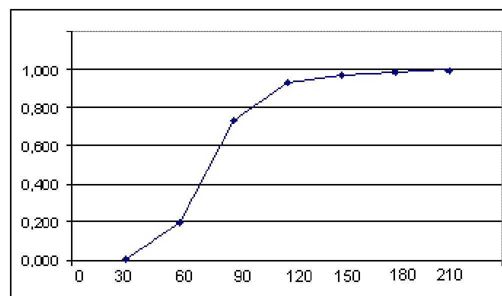
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	10.00	16	39.0	39.0	39.0
	20.00	11	26.8	26.8	65.9
	30.00	4	9.8	9.8	75.6
	40.00	6	14.6	14.6	90.2
	50.00	2	4.9	4.9	95.1
	60.00	2	4.9	4.9	100.0
	Total	41	100.0	100.0	



3.

Superfície (m^2)	ni	Ni	fi	Fi
(0 -30]	4050	4050	0,005	0,005
(30 - 60]	153900	157950	0,190	0,195
(60 - 90]	437400	595350	0,540	0,735
(90 - 120]	162000	757350	0,200	0,935
(120 - 150]	29160	786510	0,036	0,971
(150 - 180]	12150	798660	0,015	0,986
(180 - 210]	11340	810000	0,014	1,000
Total	810000		1	





c) 810000
d) 437400
e) 157950
f) 810000-786510 = 23490
g) 54% y 19,5%
h) No
i) 90
j) Si, 214650 y 26,5%
k) (60,90]
l) 120 m2
m) 150 m2
n) Puede ser.

4.

L_{i-1}-L_i	X_i	n_i	f_i	N_i	F_i
0-2	1	5	0,1	5	0,1
2-4	3	11	0,22	16	0,32
4-6	5	19	0,38	35	0,7
6-8	7	10	0,2	45	0,9
8-10	9	5	0,1	50	1
		50	1		

a) 50
b) 68%
c) 6
d) No

5.

a)

Xmin=655 Xmax=6982 n=50 número de tallos (o intervalos) como mínimo √50≈7

Unidades de los tallos: 1000

Unidades de la hoja: 100

0 | 6: representa 600

```
6      0 | 678999
24     1 | 022333444455666779
(11)   2 | 00244577999
15     3 | 03469
10     4 | 003566
4      5 | 35
2      6 | 29
```

b)

```
1 | 2: represents 1200
 leaf unit: 100
            n: 50
  6    0. | 678999
 16    1* | 0223334444
 24    1. | 55666779
 (5)   2* | 00244
 21    2. | 577999
 15    3* | 034
 12    3. | 69
 10    4* | 003
  7    4. | 566
  4    5* | 3
  3    5. | 5
  2    6* | 2
  1    6. | 9
```

c) 1400

d) 4300

6.

a) 47

b) Mín=12 Máx=82. Los valores más frecuentes (Mo) son: 33, 36 y 43.

c) 27 d) 31 e) 60

TEMA 3

1.

a) 120 días
b) $\overline{X}$=15,458
c) $\overline{X}$(10%)=15,231

2. 149,5 €/vendedor

3. a) $\overline{X}_A$= 6,4536% $\overline{X}_B$= 6% b) 6,2%

4. A) 1,4309 Euros/Libra B) 1,4276 Euros/Libras

5. $\overline{X}$=71,408

6. 19875 u.m.

7. 70%

8. 12 horas en el turno de día y 8 en el de noche

9. 8 puntos

10.

	Ej 1	Ej 2	Ej 3
$\overline{X}$	4,98	23,41	80,19
Me	5	19	76,94
Mo	5 y 6	9 y 13	(60;90)

11.

a) 0 y 300 €
b) 101 €
c) 29
d) 68 y 151 €
e) 139,5 €
f) 3,3%
g) 45 €

12.

a) Q_1=10 €
b) Q_3 = 13,8 €
c) RIQ = 3,8 €
d) D_4 = 12 €
e) C_{85} = 14,2 €

f) Frecuencia relativa 36,2%

13.

a) $\overline{X}_A$=3,35 $\overline{X}_T$= 3,40 $\overline{X}_G$= 3,39
b) 2 años
c) 4 años
d) 4 años

TEMA 4

1.

	Ej 1	Ej 2	Ej 3
$\overline{X}$	4,98	23,41	15,46
S	1,6	14,93	6,78
CV	0,32	0,64	0,44

2.

a) $\overline{X}_A$= 74,13 $\overline{X}_B$= 113,38 $\overline{X}_G$= 93,75
b) S_A= 34,17 S_B= 34,95
c) CV_A = 0,46 CV_B = 0,308 En B

3.

a) $\overline{X}$= 6,253
b) CV_A=0,279 CV_B =0,259 CV_C=0,241 CV_D=0,325. El instituto C
c) X'_D=3,5482+0,5128X_D

4.

a) $\overline{X}_A$= 630 $\overline{X}_B$= 583 $\overline{X}_C$= 632 $\overline{X}_D$= 630
b) Δ M_A = 40.000 € Δ M_B= 16.500 € Δ M_C = 41.000 € Δ M_D = 40.000 €
c) CV_A= 47,62% CV_B= 54,54% CV_C= 49,36% CV_D =50%

5.

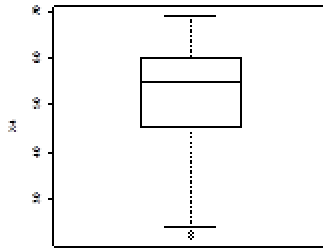
a) El empleado de la agencia 1
b) Aproximadamente 3 clientes

6. Tarde

7.

a) Barcelones, Tarragones
b) No
c) X_G= 1812,61 X_B= 3993,46 X_T= 3402,52 X_{BLl}= 1066,9 X_{VA}= 1489

8.



9. X2, X1 y X3

10.

a) $\overline{X}_R$= 11,5 mn $\overline{X}_T$= 9,2 mn
b) 15 mn
c) Z_A= 0,757 Z_B= 0,645 En la compañía A
d) CV_A = 0,83 CV_B = 0,77 En la compañía A
e) $\overline{X}_R$= 10,925 mn S= 6,51 $\overline{X}_T$= 8,74 mn

TEMA 5

1.

	H	M	Total
16-19	1	1	2
20-24	2	1	3
25-54	13	9	22
más de 55	2	1	3
Total	18	12	30

2.

a) 2,3%
b) 8,55%
c) 2,3%
d) 36,5%
e) 11,9%
f) 24,7%
g) Pasear
h) Mayores de 65
i) Aproximadamente 46 años
j) Tareas del hogar/cocinar
k) Deportes.
l) Mayores de 65

3.

a) No son independientes ya que, por ejemplo, f(UP, Esp) ≠ f(UP) f(Esp)

b) 6,5%
c) 10,6%
d)

	E	F	I	P	G	n(X)
UP	22.6	22.6	22.6	22.6	22.6	113
ASN	52.0	52.0	52.0	52.0	52.0	260
UAN	5.2	5.2	5.2	5.2	5.2	26
DAN	20.2	20.2	20.2	20.2	20.2	101
n(Y)	100	100	100	100	100	500

4.

a) 0,15
b) 0,15
c) 100
d) 60
e) 24
f) 15
g) 1
h) 1
i) 0,0375
j) 0,25
k) 0,30
l) M4

5.

a)

X \ Y	1	2	3	Total
0	0	0	1	1
1	5	0	0	5
2	10	0	0	10
3	0	9	0	9
Total	15	9	1	25

b) 52 unidades de A y 36 unidades de B

TEMA 6

1.

a) S_{XY} = 5,52 r= 0,953
b) S_{XY} = 5,80 r= 0,953

2.

a) $\begin{bmatrix} 2,08 & 1,44 \end{bmatrix}, \begin{bmatrix} 0,7433 & 0,1717 \\ & 0,34 \end{bmatrix}$

b) $r_{XY} = 0,341$

c) $S_{XY} = 0,6868$ $r_{XY} = 0,341$

3.

A r=0,32	B r= -1	C r=0,98
D r=0,03	E r=-0,42	F r=-0,95

4.

a) 0,892

b) Y*= -0,2574 + 0,3887 X

c) 79,6%

5.

a) 0,89

b) Presión*= 7,949+0,1166·Edad

c) $R^2 = 0,795$

d) Presión (m 51) = 13,896

6.

a) Y*= 14,092 - 0,00634 X

b) 58,25%

c) Se reduce 0,0634 grados

d) 8,069 grados

7.

a) 0,411

b) $X_2^* = 112,5+0,75X_1$

c) 83%

d) 937,5

8.

a) Y*= 11,5-0,75 X

b) 0,83

c) Un decremento de 0,75 unidades

d) 4,75 unidades

10.

a) $r_{(Preu.de.venda;\ Metres.quadrats)} = 0,951 < r_{(Preu.de.venda;\ Preu.inicial)} = 0,9828$.
$R^2_{II} = 0,966$; el 96,6%

b) 953,4€

c) 139821,5€

d) 254229,5€
e) la del apartado c)

TEMA 7

1.

a) E={(1,2) (1,3) (1,4) (1,5) (1,6) (1,7) (2,1) (2,3) (2,4) .... (7,1) (7,2) (7,3) (7,4) (7,5) (7,6)} En total 42 resultados elementales si se tiene en cuenta el orden de aparición.; o 21 resultados elementales si no se tiene en cuenta el orden.
b) E = { 1, 2, 3, 4, ...} Infinitos resultados numerables
c) E = $\{x \in \Re \ / \ 0 \le x \le 3\}$ Reales entre 0 y 3. Infinitos resultados continuos.
d) E = $\{x \in \Re \ / \ 0 \le x \le \infty\}$ Reales no negativos
e) E = { (0,5), (1,4), (2,3), (3,2), (4,1), (5,0)}

2.

a) P(Éxito)=3/4
b) P(copa) =1/4
c) P(Superior a 7) = 15/36
d) P(Accidente) = 1/200
e) P(llueva) = (subjetiva)
f) P(retraso) = 1/40
g) P(primo) = 1/2
h) P(3,2) =1/6

3.

a) No. $P(A \cap B)=0,6 \neq 0$
b) $P(\bar{A}) = 1 - PA) = 1 - 0,62 = 0,38$
c) $P(A \cup B)=P(A)+P(B)-P(A \cap B) = 0,62 + 0,92 - 0,60 = 0,94$
d) $P(B \cup C)=P(B)+P(C)-P(B \cap C) = 0,92 + 0,11 - 0,11 = 0,92$
e) $P(\bar{A} \cap \bar{B})=1-P(A \cup B) = 1 - 0,94 = 0,06$
f) $P(A \cup B \cup C)=P(A)+P(B)+P(C)-P(A \cap B)-P(A \cap C)-P(B \cap C)+P(A \cap B \cap C) =$
$= 0,62 + 0,92 + 0,11 - 0,60 - 0,06 - 0,11 + 0,06 = 0,94$
g) $P(\bar{A} \cap \bar{B} \cap \bar{C})=1-P(A \cup B \cup C) = 1 - 0,94 = 0,06$
h) $P(B \cap \bar{C})=P(B)-P(B \cap C) = 0,92 - 0,11=0,81$

4.

a) 2/15
b) 3/11
c) 4/15

5.

a)

	Blanco (B)	Color (C)	
Si (A)	0,05	0,10	0,15
No (A^C)	0,75	0,10	0,85
	0,80	0,20	1

b) P(A/B)=0,05/0,80 = 0,0625
c) P(C/A)=0,10/0,15 = 0,667
d) $P(A^C/C)$=0,10/0,20 = 0,5

6.

a) 0,15
b) 0,10
c) 0,39
d) 0,36
e) 0,54
f) 0,46
g) 0,2778
h) 0,2174

7. 0,016

8.

a) 0,18
b) 0,42

9.

a) 0.52
b) No $P(A\cap B) \neq 0$
c) Si $P(A\cap B) = P(A)\ P(B)$

10.

a) No. $P(B\cap C)$=0,11 P(B)P(C)=0,1012
b) Sí. P(B) = 0,92 y P(B/A) = 0,9677
c) $P(A/\ B\cap C)$ = 0,5454
d) P(A/B) = 0,652
e) $P(A/\ A\cup B\cup C)$ = 0,6595

11.

a) 0,72
b) 0,02
c) 0,26

12. 0,6

13. Del tipo V1 o V2 con probabilidad 0,375

14.
a) 0,625
b) 0,375
c) 0,4; 0,3; 0,2; 0,1
d) 0; 0,167; 0,333; 0,5

15. 0,346

16. 0,25

17. 0,779

18. 0,4167

19.
a) 0,0588
b) 0,9412
c) 0,6734
d) 0,9996
e) 0,9804

20. 0,41

21. 1/3
22. 0,818

23. 1/3

TEMA 8

1.

X	0	1	2	3
P(X)	0,216	0,432	0,288	0,064

2.
a)

X	1	2	3	4
P(X)	0,4	0,3	0,2	0,1

b)

X	1	2	3	4	5	6	7	8	...
P(X)	0,4	0,24	0,144	0,086	0,052	0,031	0,018	0,011	...

$P(x)=0,6^{x-1}0,4 \quad x=1,2,3,4...$

3.

X	1	2	3	4	5
P(X)	0,5	$0,5^2$	$0,5^3$	$0,5^4$	$0,5^4$

4.

a) 0,99
b) 0,13
c) 0,6
d) 0,333

5.

a) k=12/25 =0,48
b)

$$F(x)=\begin{cases} 0 & x<1 \\ 0,48 & 1 \le x<2 \\ 0,72 & 2 \le x<3 \\ 0,88 & 3 \le x<4 \\ 1 & x \ge 4 \end{cases}$$

c) $P(1 \le X<3)=0,72 \quad P(2<X \le 4)=0,28 \quad P(X>2)=0,28 \quad P(X<2/X<4)=0,545$

6.

a) $\sum_{\forall x} P(x)=1$

b)

$$F(x)=\begin{cases} 0 & x<1 \\ 0,05 & 1 \le x<2 \\ 0,15 & 2 \le x<3 \\ 0,30 & 3 \le x<4 \\ 0,50 & 4 \le x<5 \\ 0,70 & 5 \le x<6 \\ 0,85 & 6 \le x<7 \\ 0,95 & 7 \le x<8 \\ 1 & x \ge 8 \end{cases}$$

c) $P(X<4)=0,3 \quad P(1<X \le 4)=0,45 \quad P(X>2)=0,85 \quad P(3 \le X<6)=0,55$
$P(X \ge 3/X<7)=0,82$

7.

a) 0,0387
b) 0,3439
c) 0,81
d) 0,27

8.

a) Discreta
b)

X	1/8	1/4	3/8
P(X)	0,2	0,7	0,1

9.
b) 0,0703
c)

$$F(x)=\begin{cases} 0 & x<-1 \\ \dfrac{x^3+1}{2} & -1 \le x \le 1 \\ 1 & x>1 \end{cases}$$

10.

a) 4
b)

$$F(x)=\begin{cases} 0 & x<1 \\ \dfrac{x-1}{4} & 1 \le x \le 5 \\ 1 & x>5 \end{cases}$$

c) $P(X\le3)=0{,}5$, $P(1{,}5\le X<4)=0{,}625$, $P(X>2)=0{,}75$, $P(X>3/X>2)=0{,}667$

11.

a) 3
b) 0,992

12.

a) A- Discreta, B-Continua
b) A- $P(X\le2)=0{,}3$ $P(1\le X<3)=0{,}2$ $P(X>2)=0{,}7$ $P(X>1/X<2{,}5)=0{,}667$
B- $P(X\le2)=0{,}125$ $P(1\le X<3)=0{,}406$, $P(X>2)=0{,}875$
$P(X>1/X<2{,}5)=0{,}936$

13.

a) 0,55
b) 0,208
c) 0

d) 0,378
e) 0,552

14.
a) k=1/2
b) 0,4375
c) 71,09%
d) 0,125
e) 10,7368 Kg

15.
a) 0,5
b) 0,05
c) 0,25
d) No. CV(X) < CV(Y)

16.
a) 0,25
b) 0,2727

X	3	5	7	8
P(X)	1/12	3/12	5/12	3/12

a) Me=7, E(X)=6,417 D(X)=1,5

17. E(X1) = 650   E(X2)=250   E(X3)=400
V(X1) = 15427500 V(X2)=562500 V(X3)=0
CV(X1) = 604% CV(X2)=300%

18. E(Coste)=273

19.
a) 0,75
b) 0,625
c) E(X)=49,8 V(X)=1,66
d) E(C)=114,6 V(C)= 6,64

20. E(X) = 0 V(X) = 0,6 D(X) = 0,774 (ejercicio 9)
E(X) = 2,25 V(X) = 0,3375 D(X) = 0,58 (ejercicio 11)
$E(X_A)$ = 2,9 V(X) = 1,49 D(X) = 1,22 (ejercicio 12)
$E(X_B)$ = 3 V(X) = 0,6 D(X) = 0,774 (ejercicio 12)

21.

a) k= 3
b) E(X) = 1,5 Me = 2 Mo = 4
c) V(X) = 6,25 D(X) = 2,5
d)

Z	-1,4	-0,2	0,2	1
P(Z)	0,3	0,1	0,2	0,4

22. 0€

23. 5€

24.

a) 2,44
b) 2,26
c) 3,36
d) V(X) = 1,13 D(X) = 1,06
e) Si

TEMA 9

1.

a) 0,311
b) 0,5
c) 0,5
d) 0,6329
e) 0,1875
f) 0,4744

2.

a) 0,1240
b) 0,9743
c) 0,1497

3.

a) 0,5987
b) 0,086
c) 0,9139
d) 0
e) 0,7852
f) 0,5993
g) 0
h) 0,2578

4.

a) $E(X)=5$ $D(X)=1,94$
b) $E(Y)=1,25$ $D(Y)=2,42$

5.

a) 0,1160
b) 0,2373
c) 0,0275

6.

a) 0,0755
b) $V(X) =0,95$ $D(X)=0,975$
c) 0,8286
d) 0,08614

7.

a) 0,4752
b) 0,1792
c) $E(X)=800$ $D(X)=489,9$

8. 0,9162

9.

a) 0,9772
b) 0,1922
c) 0,3086
d) 0,835
e) 0,9544
f) 0,5154

10.

a) 65,3
b) 11
c) 216,5
d) 6,8

11.
a) 61,79%
b) 8,95
c) 23
d) 23,1 puntos

12.
a) 0,15866
b) 0,31084
c) 617
d) 0,88493
e) Inferior a 448 gr.
f) 0,901

13. 166500 litros

14.
a) 0,04746
b) 4,992 mn

15. 71,6%

FORMULARI0

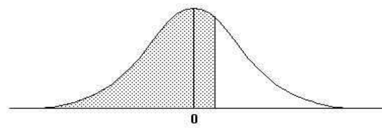
$$Me = L_{i-1} + \frac{0{,}5n - N_{i-1}}{n_i} a_i \qquad s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$$

$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{n-1} \qquad r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

$$\hat{Y}_i = a + bX_i \qquad b = \frac{S_{XY}}{S_X^2} \qquad a = \bar{Y} - b\bar{X}$$

$$\mu = E(X) = \begin{cases} \sum_{\forall x} x P(x) \\ \int_{-\infty}^{+\infty} x f(x)\,dx \end{cases} \qquad \sigma^2 = V(X) = E(X-\mu)^2 = \begin{cases} \sum_{\forall x} x^2 P(x) - \mu^2 \\ \int_{-\infty}^{+\infty} x^2 f(x)\,dx - \mu^2 \end{cases}$$

5. FUNCIÓ DE DISTRIBUCIÓ NORMAL TIPIFICADA



$$F(z)= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t/2} \, dt$$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

BIBLIOGRAFIA

ALEA, V., JIMÉNEZ, E., MUÑOZ, MC., TORRELLES; E., VILADOMIU, N. (2014) Guía para el análisis estadístico con R-Commander. Barcelona, Text docent 391 UB

ALEA, V., MAQUEDA, I., MUÑOZ, MC., TORRELLES; E., VILADOMIU, N. (2011) Estadística I. Cuestiones tipo test con R-Commander. Barcelona, Text docent 368 UB

ALEA, V., MAQUEDA, I., MUÑOZ, MC., TORRELLES; E., VILADOMIU, N. (2001) Estadística para las ciencias Sociales: Cuestiones tipo test Madrid, Thomson

ALEA, V., MAQUEDA, I., MUÑOZ, MC., TORRELLES; E., VILADOMIU, N. (2009) Introducción al análisis estadístico con R-Commander. Barcelona, UB

LIND, A.D. et. al. (2008) Estadística aplicada a los negocios y la economía. Madrid, Mc Graw-Hill

MARTÍN-GUZMÁN, P (2006) Manual de Estadística: Descriptiva. Madrid, Thomson

MONTIEL, A.M. et. al. (1997) Elementos básicos de Estadística Económica y Empresarial. Madrid, Prentice-Hall

NEWBOLD, P. et. al. (1997) Estadística para los negocios y la economía. Madrid, Prentice Hall

PEÑA, D., ROMO, J. (1997) Introducción a la estadística para las ciencias sociales. Madrid, McGraw-Hill